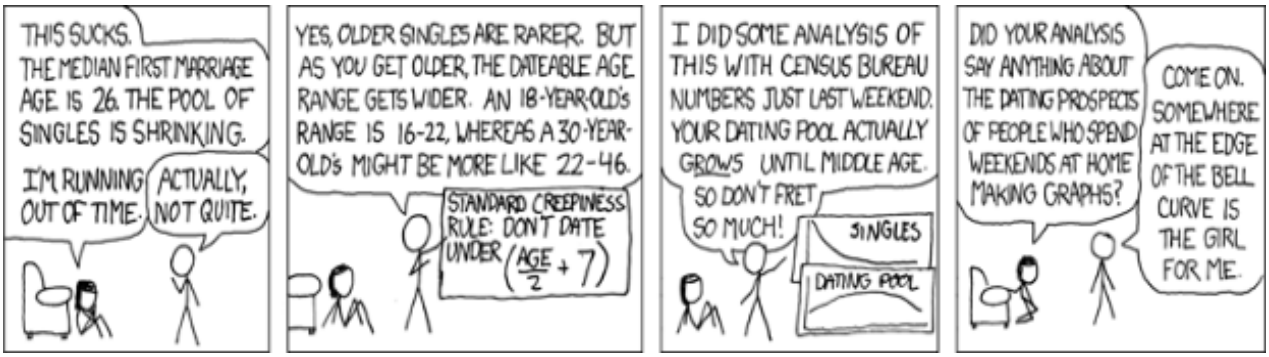
Chapter 1: Introduction to Statistics



1.1 You've Decided to take Statistics

How well is our government performing? At the time of my revising this book (June 2015), the United States Congress is enjoying a 19% approval rating and 77% disapproval rating, President Barack Obama has a 47% approval rating and a 48% disapproval rating, and the US Supreme Court has a 44% approval rating and a 48% dispaproval rating (www.gallup.com; Liptak, 2012). In essence, two-thirds of our government (the executive and judicial branches) are doing a 'so-so' job, but the other one-third (legislative branch) is doing a lousy job. Interestingly, nearly 90% of the current members of congress will be re-elected.

Only a few years after the 200th anniversary of Charles Darwin's birth and a little over 150 years after publication of Origin of Species, 42% of Americans still believe in creationism and surveys show only 4 in 10 Americans believe Evolution, 2 in 10 reject Evolution in favor of a biblical interpretation of the origin of human life on earth (e.g., creationism, intelligent design), and 4 in 10 are unsure on the issue. Although this trend is influenced by education level (Newport, 2009), this is odd given there is an abundance of empirical evidence that supports evolution.

At the end of the 2014 Major League Baseball season, the New York Yankees had a winning percentage of 51.9 (proportion of games won =.519), which was not enough to win the American League East Division.

Contrary to popular myth, there is no relationship between the shoe/foot size and penis length (Shah & Christopher, 2002). However, sexual intercourse without a condom is associated with fewer depressive symptoms in women (Gallup, Burch & Platek, 2002).

What do these have in common? They involve statistics. If statistics had not been invented we could complain about congress, the president, and the Supreme Court, but could not describe by *how much* we generally approve or disapprove of their performances. If we did not have statistics, we would not know *how many* people accept evolution (or don't). If we did not have statistics, we could not determine *how much better* one sports team performed relative to other teams to determine playoff standings. Finally, without statistics, we could not *disconfirm* myths about the relationship between the sizes of various body parts.

Statistics are used just about everywhere and in most occupations. Hence, understanding statistics, including where they are used, why they are used, what they represent, how to interpret them, and how to properly use them is an important part of being an informed citizen. Understanding statistics allows you to understand facts and figures presented in the media and in other college courses and allows you to be a little more skeptical about information when it is presented. If you do not know how to question information, it is unlikely that you ever will.

Unfortunately a lot of students are scared of statistics and research-based courses, because they hear horror stories from former students. Here are a few statements from my conversations with former students: "*The class is horrible, boring, and stupid!*"; "*You cannot get an A in the class, ever*."; "*The material is terrible, dry and evil*."," *Please kill me.*", "*Kill me.*", "*I want mommy!*" I may have exaggerated those last few. Let me dispel some of these comments:

First, it's true a professor or a course can be boring and make something more difficult than it should be, but ultimately your education is what you make of it. If you don't like something, *speak up*! But, if something is truly horrible, boring, and stupid, perhaps this course or your major is not right for you (just saying).

Second, I'll admit statistics is a bit dry and the material can be boring, which is why most examples in class and this book involve real life, society, elves, wolverines, Bigfoot, zombies, the media, poking fun at the ruling class, and TV shows (I've gotta' have fun with my job too!) But make up your own examples if you don't like mine. What works for me may not work for you and that's okay. Make it fun. It's just statistics!

Third, it is possible to do well in this course. Any student that has received a grade of C- or less from me in statistics did so by their own hand, by not seeking help when needed. Remember, I am your *guide*: use me! By the way, don't worry about your grade: Worry about learning and comprehending the material; your grade will fall into place.

Finally, the material in statistics is difficult, but is doable and you can earn a good grade. Statistics uses basic mathematical principles (addition, subtraction, multiplication, division, etc), but uses them in different ways than you may be used to, so some students come into to a statistics course with preconceptions of it being impossible. Nothing is impossible...except making the Kessel Run in less than 12 parsecs.

1.2 What, When, Where, How, and Why of Statistics

Below, are answers to the basic what, when, and where questions that are related to statistics:

What are Statistics? Statistics are procedures that combine, organize, and summarize data to make inferences. Scientists collect data to generate and test theories and hypotheses, but the data are initially unorganized and 'raw', and statistics allow one to organize the raw data into something condensed and more meaningful. Statistical methods also allow one to examine and quantify relationships among variables to uncover trends, and to determine whether a theory or hypothesis is supported by that data.

When are Statistics Used? Statistics are used whenever you are trying to answer a question that requires summarizing data. Whether it is a psychological study on political attitudes and personality or a newspaper survey of how much readers like a politician. Whenever you have a question that requires a summary of raw data, you'll use statistics to answer the question.

For example: Let's say I hypothesize that age is related to political attitude (liberal-conservative); specifically, older individuals tend to be more conservative. One way to address this would be to ask people their age and their political attitude on a liberal-conservative scale. Then, based on the data I collect, I can determine whether there is a relationship between age and political attitude. A set of hypothetical data listing ages and political attitudes of people is presented below (Political Attitude was measured using an 11-point scale, from 1 = liberal to 11 = conservative). You can see that as age increases, political attitudes tend to become more conservative. Although the data appear to support my hypothesis, the only way to know whether a relationship exists is through statistical analysis.

Age	Political Attitude
15	3.0
20	3.5
25	5.0
30	5.5

35	6.0
40	6.5
45	8.0
55	9.0
60	9.5

Where are Statistics Found? Statistics are not only used by scientists; all professions use them. Businesses, teachers, schools, sports, government, the military, the justice system, religions, and social organizations use statistics in their own ways. For example, at the end of the 2007 Major League Baseball season the batting average (AVG) of Derek Jeter was .322, which means he hit a pitched ball and made it to safely to base about 32.2% of the times he made a plate appearance. This value is a type of statistic that players, teams, sports writers, and fans use to gauge how well a player is performing.

As another example, at the time of my revising this book, President Obama is enjoying a 48% approval rating with a margin of error of ±3%. This survey, performed by Gallup, included about 1,500 adults. People use these numbers, based off of a small group of people to infer what percentage of all Americans would approve or disapprove of President Obama's job performance. Indeed, statistics uses data from small groups or *samples* to makes inferences about what would be found in a larger group or *population*.

Why do I need Statistics? There are a lot of reasons, here are a few: First, You need to understand statistics in order to understand most scientific research. If you take a research methods course in the future, your professor will discuss use a lot of data analysis examples. Second, statistics are used to explain topics in upper-level psychology courses, especially for more empirically-based courses like Sensation and Perception, Cognition, Conditioning and Learning, Social Psychology, and courses in Neuroscience. Lastly, statistics allows you to understand information presented in the media. You always hear percentages and proportions of people that agree or disagree with some topic. For example, in President Obama's approval ratings from above, I mentioned that the margin of error was ±3% in that survey, but what does this mean? If you know something about statistics you will be able to understand. Also, knowing statistics allows you to judge for yourself whether something in the popular media makes sense or not.

1.3 Statistics as a Language

Dr. Burnham's First Law of Statistics: ***Statistics is a language course more than a math course.***

Dr. Burnham's Second Law of Statistics: ***Repeat the First Law.***

Dr. Burnham's Third Law of Statistics: ***Don't be a dumbass.***

I know what you are thinking, what is he talking about? The chapters in this book are filled with numbers and a lot of equations, how is this *not* a math course? How can statistics be a language course?

It's true statistics involves numbers and involves using formulas to solve problems. But, there is more to statistics than numbers and formulas and *this* should be your focus in the course. The concepts, content, and *language* of statistics is key to understanding how to use the formulas appropriately and how to properly interpret statistical information.
Why is learning statistics like learning a language? Statistics is best understood at a conceptual level just like any language. When you learning a new language you don't focus only on the new words in isolation, you relate the new words to the words you know in your native language. Similarly, the mathematical operations used in statistics are hose you should already be familiar with, including adding, subtracting, multiplying, and dividing. The difficult part to statistics is you will use these basic mathematical concepts in new and different ways and this is where students get into trouble. Students tend to focus solely on each term in each formula, but fail to integrate concepts to comprehend the overall meaning statistical procedures. For example, when you read a word (e.g., corruption) and try to retrieve the meaning of that word from your mental dictionary, you don't focus on each letter in the word (c-o-r-r-u-p-t-i-o-n) to retrieve

its meaning; rather, you focus on the entire word. The same is true in statistics. Don't focus on the small pieces of formulas and the numbers, focus on the language and on the big picture. If you do this, the little pieces and the numbers fall into place.

Also, because you will likely be reading research reports, articles, and textbooks that include statistics you need to learn the language of statistics. For example, when you see *SD* = 12.558-sec, you need to know how to interpret this statistically. Thus, there is a language to statistics that you will learn to comprehend.

1.4 Two Types of Statistics

There are actually two types of statistics: *descriptive statistics* and *inferential statistics*. *Descriptive statistics* are used to consolidate data it into single, representative values that can be used to describe the entire set of data.

For example, take the following set of data {3, 4, 6, 7, 3, 4, 5, 8}. You might take the average of these scores (average = 5) and use the average to represent the data set. As another example, you might have the following data {7, 7, 7, 7, 7, 7, 7, 7} and use 7 as a representative value, because 7 is the only value in the data. These are types of descriptive statistics that summarize the individual values in a set of data.

Inferential statistics serve two purposes: First, they statistical tests we use to analyze relationships between variables. Second, inferential statistics allow us to make inferences and assumptions about the data we collect relative to some larger group that was want to know something about. That is, we cannot collect data on very large groups, because often don't know the size of a group, so we collect data from smaller samples and use that data to make inferences about what we would likely find in the larger group. This is using statistics to make inferences; hence, inferential statistics.

1.5 Variables

Variables are anything that can change or take on different forms. For example, blood pressure is a variable, because it can vary across people and vary across time for a single person. Sex (or gender) is a variable, because it differs across people. Political attitude is a variable, because some people are liberals and some people are conservatives. Variables are often labeled with an uppercase letter such as X or Y; for example, if I measure political attitudes, I might use X to denote the values for the variable Political Attitude.

If something has only one level or can take on only one form, it is a *constant.* Constants are anything that remains stable and does not change across time or across individuals. For example, a person's sex (their being biologically male or female) is constant, because biological sex cannot change. Numbers are constants, because 5 has the same value in any situation.

In psychology and most of the other behavioral sciences, variables are *hypothetical constructs*, that is, something we believe exists and can explain behavior, but at the same time is not directly observable. For example, *memory* is a hypothetical construct because we cannot actually see memory (we cannot see the things a person remembers), but we think memory exists, because information a person retains influences their behavior.

Operational definitions are clear and concise descriptions of the procedures used to manipulate or to measure a hypothetical construct. Because most constructs are not observable we must define an indirect method of measuring. For I might operationally define the hypothetical construct 'memory' as the 'number of words recalled from a list of 100 words studied 24 hours earlier'. I will not dwell on operational definitions, just note when I say we are measuring something unobservable like memory, I mean we are measuring something related to memory.

In statistics, *dependent variables* are anything that is being measured. Dependent variables are within the operational definition for a hypothetical construct, so if we define 'memory' as the number of words recalled from a list of 100 words studied 24 hours earlier, the number of words recalled is the dependent variable of memory. Alternatively, if we measure political attitude using a rating scale from 1 (liberal) to 11 (conservative), the rating scale is the dependent variable for political attitude. Thus, the construct is not the dependent variable; what is directly measured is the dependent variable.

Independent variables are anything that can change or differ between situations or groups and that can influence a dependent variable. Independent variables can be something that differs naturally between groups or situations, for example, biological sex (male vs. female); and independent variables can also be directly manipulated across groups or situations, for example, giving one group of people an antianxiety drug and giving another group an inert placebo (independent variables that are manipulated are often called *true independent variables*).

For example, I am interested in the influence of noise on studying. I assign one group of 10 students to study between 7:00PM to 8:00PM each night while I play loud music and a second group of 10 students to study during the same time, but without music; hence, I *manipulated* noise between the two groups. To examine the influence of noise on studying, I could compare exam scores on the studied material between groups. If the manipulation of noise influenced studying as measured by the exam scores, I might expect the results below.

Noisy	**Quiet**
70% Correct	85% Correct

You can see the change in the independent variable (noisy versus quiet) was associated with a change in the dependent variable (% correct). Before I can conclude the difference in the % correct was due to the noise, I must use inferential statistics because it is possible this difference happened by coincidence.

Extraneous variables are unrelated to the dependent variable and independent variable and are allowed to randomly vary. An example of an extraneous variable might be a person's age. It is unlikely every person in a study has the same birthday, so age is allowed to vary randomly unless it is a variable that should be controlled. A *control variable*, which is actually an extraneous variable that is made into a constant, is something that is extraneous to a study, but remains constant across the levels of independent variable. For example, studies of attention often include only right-handed individuals; hence, handedness is a control variable.

Extraneous variables are not a problem unless they become *confounding variables*, which are unrelated to a study, but change as levels of the independent variable change. From the example above, say students in the noisy condition had an average GPA of 2.00 while the average GPA for students in the quiet condition was 3.67. In this case the manipulation of noisy versus quiet studying conditions was confounded with GPA. (See table below) The difference in GPA across levels of the independent variable is a problem, because there is an alternative explanation for the observed difference in % correct: It could be the higher % correct in the quiet condition was due to that group having smarter students.

	Noisy	**Quiet**
Exam Grade	70%	85%
Average GPA	2.00	3.67

1.6 Values and Levels vs. Variables

You need to careful when referring to variables so you don't refer them by their values or levels, that is, be sure you don't confuse a variable with one of the possible levels or values it can take on. For example, a study on college students may require subjects to identify themselves by *college class* and each subject

would respond whether they are a *freshman*, *sophomore*, *junior*, or *senior*. These are the *values* or *levels* of the variable 'college class', thus, 'senior' and 'sophomore' are not the variable, they are two of the levels.

1.7 Quantity vs. Quality

One difference among variables is whether they are quantitative or qualitative. A *quantitative variable* includes numbers as levels. Thus, a quantitative variable is one where levels of the variable are numerical and can be used to express differences in amounts or ratios. Examples include height, age, GPA, temperature, salary, length, etc.

A *qualitative variable* (or *categorical variable*) is not quantitative and the levels include distinct categories. Thus, qualitative variables do not have numerical meanings attached to their levels, so data collected from qualitative variables cannot express differences in amounts or ratios, but can express differences in types. Examples include biological sex (male vs. female), hair color, college major, college class, race, political orientation, national affiliation, etc. Qualitative variables may be disguised as quantitative variables. One example is in digital data files (e.g., Excel, SPSS, R), where the levels of the variable 'sex' may be entered as 1 for male and 0 for female. Assigning numbers to categories is called *dummy-coding* and is done because it is easier to work with and to manipulate numbers in data files than it is to manipulate categories. In this case 'sex' is still qualitative, because 'sex' is not actually measured as a number.

Another example of a qualitative variable being disguised as a quantitative variable is a *Likert scale*, which is a rating scale where numbers are assigned to different levels of agreement or likability. For example, here is a question one of my colleagues use in studies on political orientations:

Overall, what is your general political attitude?

-4	-3	-2	-1	0	+1	+2	+3	+4
very liberal	moderately liberal	somewhat liberal	slightly liberal	totally neutral	slightly conservative	somewhat conservative	moderately conservative	very conservative

Subjects are asked to circle the number coinciding with their level of liberalism-conservatism. The numbers used are actually meaningless, because I could use the following scale with different values, but where the categorical levels are the same:

Overall, what is your general political attitude?

1	2	3	4	5	6	7	8	9
very liberal	moderately liberal	somewhat liberal	slightly liberal	totally neutral	slightly conservative	somewhat conservative	moderately conservative	very conservative

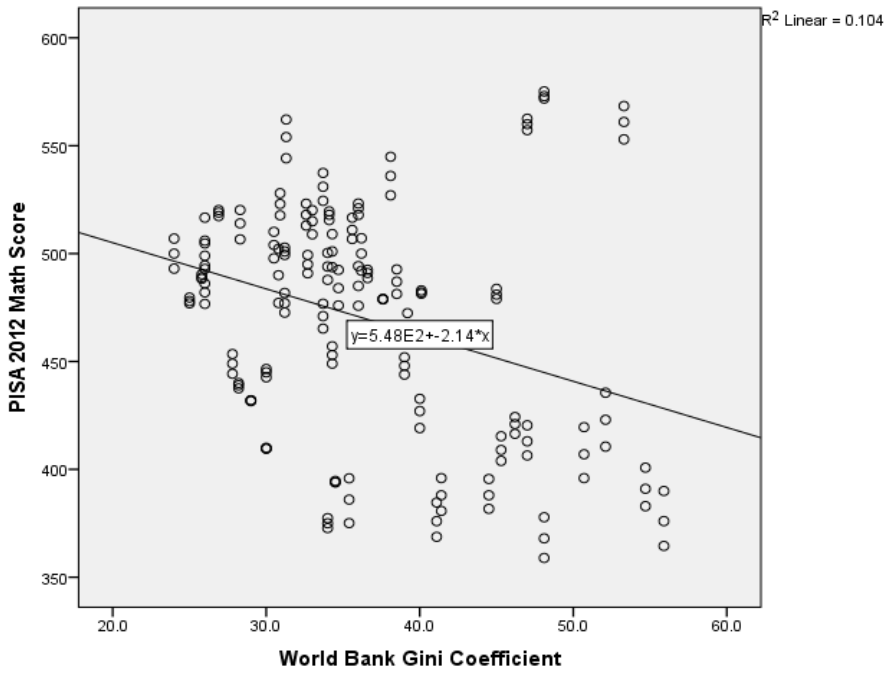
Thus, people are responding how liberal-conservative their political orientation is, which is qualitative, <u>not</u> quantitative. Indeed, say two subjects (A and B) are measure on this scale. Subject A selects a 2 (very liberal) and subject B selects 10 (very conservative). Taking the average of these two measurements I get a value of 6 [(2+10)/2=6], which is associated with 'totally neutral'. Were these two individuals neutral in their political orientation? Hardly! But based on a quantitative analysis of qualitative data I would erroneously make this conclusion. Thus, be careful when assessing and interpreting Likert-scale data. These scales are used throughout psychology and the results can be misleading.

The qualitative vs. quantitative issue is important, so you should watch for these terms throughout the book. Indeed, the types of variables (qualitative vs. quantitative) used dictates the type of inferential and descriptive statistics you use to analyze data.

1.8 Hypotheses and Predictions

Statistics are used, mainly, to evaluate research questions. When a researcher has a question, s/he designs a study, collects data, and applies appropriate statistical procedures to determine whether the data support his/her question. Everything centers on the research question, or *hypothesis*, which generates testable predictions.

Say I am a interested in the relationship between wealth inequality as measured by the Gini coefficient and performance on mathematics tests. The Gini coefficient is a measure of national wealth inequality and is calculated by many organizations around the world, where a value of 0 indicates complete equality and 1 (or 100%) indicates complete inequality. To measure mathematics performance I use the most recent Program for International Student Assessment (PISA) test scores in math.



There are two hypotheses we must generate: a *null hypothesis* (H_0) and the *alternate hypothesis* (H_1). Each includes a specific prediction about what might be found when data are collected, but the null hypothesis predicts the research question will not be supported by the data; thus, it predicts no relationship would be found between wealth inequality and mathematics test performance. In contrast, the alternate hypothesis predicts a research question will be supported by the data; thus, it predicts a relationship will be observed between wealth inequality and mathematics test performance. Hence, the two hypotheses make opposing predictions: the null predicts no difference in the dependent variable or no relationship, whereas the alternate predicts a difference in the dependent variable or a relationship. A graph of these data is on the preceding page. Do you think there is a relationship between wealth inequality and math test performance or not? If there is, does this seem to support the null hypothesis or the alternate hypothesis?

1.9 Populations and Samples

Any scientist has a target group of interest they want to study. For example, if one is interested in studying the behavioral consequences of having seasonal affective disorder, the target group would be all people who have suffered from seasonal affective disorder. On the other hand if you are interested in studying the effectiveness of a new statistics teaching method, your target group would be all statistics students.

A *population* includes all members of that target group and the number of people in that population is usually labeled with an uppercase *N.* In the first example above, the population is all people who have suffered from seasonal affective disorder and in the second example the population is all statistics students.

Researchers make inferences about populations, so they need to be sure to collect data from the correct population. For example, you might conduct research on depression, but the people with whom you conduct the research were only males between the ages of 20 and 25 with mild depression. In this case, you cannot

make inferences about all people with depression; rather, you can make inferences only about males between the ages of 20-25 with mild depression. Thus, it is critical to have research participants reflect your intended target population.

Any statistic based on a population is *parameter*. For example, the average depression of all males with mild depression would be a parameter and the batting average of all hitters in Major League Baseball would be a parameter. Generally, the symbols used to denote parameters are Greek symbols such as σ (sigma), ρ (rho), μ (myu), and ω (omega). I have listed some of the more common symbols table below.

It is often impractical and impossible to conduct research using entire populations, because most populations are infinitely large, that is, we do not know for certain how many people are in the population and the size is always changing.

Instead of collecting data from populations, researchers rely on *samples* of populations, with samples being smaller, representative groups selected from the population. As long as the sample is representative of the population, a researcher can use the data collected from the sample to make inferences about what would likely be found in the population. That is, as long as individual differences from a population are equally represented in a sample, a researcher can make conclusions about what would be found in the population based on what is found in the population.

Any statistic that is based on sample data is a *statistic*. For example, the batting average of a randomly selected group of 100 Major League Baseball players is a sample statistic. Generally the symbols used to label statistics are English letters (usually lowercase), such *r*, and *t*. Some of the more common symbols and their meanings are listed in the table below. Importantly, when you use sample data to estimate a parameter it is customary to include a ^ (hat) above the symbol for the sample statistic. Please be aware of this! Sample statistics and population estimates from samples have the same general meaning, but are calculated differently and refer to different things. Sample statistics measure something within a sample and parameter estimates from sample estimate the parameter that would be found in a population. If you want to describe only the sample, use a sample statistic, but if you want to use a sample to make inferences about the population then use a population estimate.

Some Common Parameter, Statistic, and PArameter Estimate Symbols.

	Measure of...	Population Parameter	Sample Statistic	Parameter Estimate
Mean	Arithmetic average	μ	$\overline{X}$ or *M*	$\overline{X}$ or *M*
Population Size	Number of subjects	N	n	n
Variance	Average variability	σ^2	s^2	$\hat{?}^2$
Standard Deviation	Average distance between a value and the mean	σ	s	$\hat{?}$
Correlation	Strength of relationship between two variables	ρ	r	r

1.10 Sampling

The main issue with samples is whether they are representative of a population, whether the subjects in the sample accurately reflect the subjects of the population. The best way to ensure a representative sample is through *random sampling*, which involves selecting people from a population at random and without bias. For example, if I were to select a random sample of baseball players for a sample, I might assign each Major League Baseball player a number and randomly select 10 of those numbers, with the 10 selected players becoming part of my sample.

The downside to sampling is occasionally and even with random sampling a sample becomes *biased* and not representative of the population. This occurs when you get too many individuals from a certain sub-

group within a population into your sample. For example, you get too many males or too many females in a sample, or too many people with a high IQ. Unfortunately, it is difficult to detect this and the only way to ensure a sample is representative is measure people in a sample on relevant variables and check.

Another issue is whether you *sample with replacement* or *sample without replacement*. When you sample with replacement, after you select someone for a sample you place that person back into the population so they could be selected again. In this case, because the population will always have the same number of individuals the chance of being selected is the same, but the same person could be selected more than once. In sampling without replacement, after someone is selected from a population and placed into a sample they are not returned to the population and cannot be re-selected. In this case, a person cannot be selected more than once, but after every selection the size of the population decreases by one person and the chance of being selected increases over time. This only becomes a problem if you have a small population. If the population is large, which is usually the case it is of little concern.

1.11 Random Selection

Say I design a study that will examine the differences in attention between video game players and non-players. The study involves a survey to assess the amount and types of video games people play and a computerized task to assess attention. I determine I need n = 50 subjects for this study and luckily I am teaching two sections of statistics, each with 25 students. I could recruit all students in these two sections for my study, but is there a problem using my 50 statistics students as my sample? Yes, there is, and the the problem is sample is being used out of convenience and is probably not representative of my target population. Actually, in this example I have two intended populations: (1) people who play video games, and (2) people who do not play video games. It is entirely possible I would find some video game players and some non-players among my 50 students, but how well do they represent the general population of video gamers and non-gamers? It's tough to say.

When creating a sample the most important thing is the sample must be representative of the population. This means that individual differences among subjects in the population are represented proportionally in the sample. What I mean by 'represented proportionally' is large sub-groups within a population have more subjects in a sample relative to smaller sub-groups within the population.

As another example, let's say that I run a study that will examine the relationship between vitamin C intake and intelligence. Assume I teach at the Massachusetts Institute of Technology (MIT) and run the study there. What's the problem with using a sample of MIT students for my study? MIT has a lot of highly intelligent individuals, so the sample would not be representative of all intelligences levels, which could skew the results. Thus, when selecting subjects for a sample out of convenience, there is a low probability that your sample will be representative of your intended population.

Aside from representativeness, there is a second problem with using my 50 statistics students for a sample. By using these subjects out of convenience I have introduced *selection bias* into my sampling procedure. This bias stems from the fact that I allowed only certain individuals into my study (e.g., college-educated, statistics students, aged 18-22, etc.). Using a sample for convenience allow individuals with only certain characteristics into a sample, which produces *biased samples*.

The solution is *random sampling*, which is simply selecting individuals for a sample in a random, unbiased manner. Think of random sampling as putting everyone's name from a population into a hat and randomly drawing a certain number of names from the hat to create the sample. Instead of putting names into a hat and drawing names at random, modern researchers generally do something a little more technologically advanced. Modern random sampling makes use of *random number generators*, which function exactly how their name implies: they generate random numbers. The random numbers that are generated represent those individuals who are selected for a sample.

For example, say I have the following population of red and blue letters X and O and I need to create a sample of *n* = 10 from this population. Assume there are equal numbers of red and blue Xs and Os in the population (*N* = 10 for each group in the population). Here's a graphic representation of this population:

X X X X X X X X X X
O O O O O O O O O O
X X X X X X X X X X
O O O O O O O O O O

The first step to random sampling is to assign each potential subject a number. In this case, I'll just use 1 – 40, which I have placed as subscripts below:

X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9 X_{10}
O_{11} O_{12} O_{13} O_{14} O_{15} O_{16} O_{17} O_{18} O_{19} O_{20}
X_{21} X_{22} X_{23} X_{24} X_{25} X_{26} X_{27} X_{28} X_{29} X_{30}
O_{31} O_{32} O_{33} O_{34} O_{35} O_{36} O_{37} O_{38} O_{39} O_{40}

Next, using a random number generator (e.g., http://www.randomizer.org/), I generate a set of 10 random numbers, because I need a sample of *n* = 10. Assume I obtain the following set of 10 randomly-generated numbers: {1, 3, 10, 16, 18, 21, 22, 34, 38, and 40}. Subjects with these numbers are placed into my sample:

X_1 X_3 X_{10} O_{16} O_{18} X_{22} X_{22} O_{34} O_{38} O_{40}

Random selection attempts to have each sub-group within a population be represented proportionally within a sample. In this case, it would mean having equal numbers of red Xs, red Os, blue Xs, and blue Os in the sample. Indeed, you can see above there are approximately equal numbers of red Xs, red Os, blue Xs and blue Os in the sample. Thus, the sample is fairly representative of the population.

Does random selection guarantee a sample will be representative? No, because it is also possible a random number generator will give the following set of ten numbers {1, 2, 3, 4, 5, 6, 7, 8, 9, and 10}. This is just as likely to happen as the set above. Indeed, a biased sample is one possible randomly generated sample. The point of random assignment is to have individual differences in a population distributed proportionally in a sample. If random selection is effective, larger groups in a population will have more members in a sample and smaller groups within a population will have fewer members in a sample.

1.12 Random Assignment

In a particular population, say there are happy people (☺), frowny people (☞), and non-emotional people (☠). Through random selection you obtain the following sample of *n* = 12 people, which contains ten non-emotional people and two frowny people:

N N N N F N
N N N F N N

You need to create two groups from the *n* = 12 subjects in your sample. You could simply split the group in half, where the subjects on the left are placed into Group A and the subjects on the right are placed into Group B:

Group A			Group B		
☠	N	N	☠	☞	☠
N	N	N	☞	☠	☠

Unfortunately, group B has the two frowny individuals from the sample and group A has only non-emotional individuals. What's the problem? Say you manipulate an independent variable between groups A and B and measure some dependent variable in each group. For example, individuals in each group are timed for long it takes to solve a puzzle, but for group A the puzzle is easy and for group B the puzzle is hard. The average time for individuals in group A to solve the easy puzzle is 2 minutes and the average time for the individuals in group B to solve the hard puzzle is 4 minutes. In this case, it would appear that solving hard puzzles requires more time than solving easy puzzles, but there is a problem: the two frowny individuals in group B make group B naturally different from group A; hence, the difference in completion times could be attributed to this fundamental difference in the makeup of the groups. Perhaps frowny people take longer time to solve puzzles and made the overall average completion time longer than would be expected.

What's the solution? First, it is extremely unlikely you would know for certain that you have two frowny individuals and 10 non-emotional individuals in your sample and that's the catch. The only way to know what individuals are in a sample is to assess them and their personality, mood, blood type, favorite Kafka quote, etc. But, researchers don't do this, because we don't have the time or the need unless frowny-ness is relevant. The solution to this problem above is *random assignment*, which is placing subjects from a sample into different groups at random and without bias. Random assignment is sort of like random selection, where everyone in the sample is assigned a number and then numbers/subjects are randomly selected to be placed into group A or group B. For example, say that the *n* = 12 individuals from above are assigned the following numbers:

☠ = 1 ☠ = 2 ☠ = 3 ☠ = 4 ☞ = 5 ☠ = 6
☠ = 7 ☠ = 8 ☠ = 9 ☞ = 10 ☠ = 11 ☞ = 12

Using a random number generator (http://www.randomizer.org/), I generate a set of six random numbers {2, 4, 5, 7, 11, and 12}. Individuals with these numbers will be placed into group A and the other subjects will be group B:

Group A			Group B		
☠ = 2	☠ = 4	☞ = 5	☠ = 1	☠ = 3	☠ = 6
☠ = 7	☠ = 11	☠ = 12	☠ = 8	☠ = 9	☞ = 10

Now, you have one frowny individual in each group so there is an equal distribution of frowny people and non-emotional people between groups. The point of random assignment is to increase the chance that individual differences in your sample are distributed equally between groups. If random assignment is effective, any difference between groups is likely due to the manipulation of an independent variable and not due to a priori differences between the groups. However, there are two important points to make:

First, random assignment does not ensure individual differences in a sample will be distributed equally between groups; random assignment only increases the likelihood of this happening. For example, another set of random numbers might be {2, 4, 5, 8, 10, and 11}. In this case, both frowny individuals (underlined numbers) would be in group A; hence, there is no guarantee random assignment will work.

Second, there is theoretically an infinite number of individual differences across people; it's what makes everybody, usually, pretty cool. We cannot account for every single one of them; hence, it is likely there will be subtle differences between groups based on individual characteristics. This is not meant to make random assignment seem futile or unnecessary; the best chance we have of equally distributing all of those individual differences, known and unknown, is by randomly assigning people to groups.

1.13 “Law” of Large Numbers

In later chapters I discuss such topics as *effect size*, *statistical power*, *directionality*, and *alpha level*. All of these are relevant to determining the appropriate sample size that should be used for a given research study, but right now it is premature to address this. So, please keep in mind this section is going to be discussed in very general terms.

There is something to be said for *strength in numbers*; this is why The Borg is a relentless force in the Star Trek universe. Statistically speaking larger samples are generally better (*strength in numbers*). This issue will come up repeatedly throughout this book and the class so please keep this in mind. Most of the reasons for larger sample sizes are quantitative in nature and we’ll deal with those as they arise. Below, I lay out a simple case of why larger samples tend to be more representative of a population (*strength in numbers*).

Say I have the following population (*N* = 30) of frowny, happy, and stoic people:

☠ ☠ ☠ ☠ ☠ ☠ ☠ ☠ ☠ ☠ ☠ ☠
☠ ☠ ☠ ☠ ☠ ☠ ☠ ☠ ☠ ☠ ☠ ☠
☞ ☞ ☞ ☟ ☟ ☟

Most people in this population are stoics and only a few people are happy or frowny. We call these individuals who occur infrequently in a population ‘unique’ or *outliers*. The outliers, though part of the population and should be part of the sample, are definitely not the most representative of the overall population. Hence, you do not want many of them in a sample. I randomly select a sample of *n* = 5:

☟ ☟ ☠ ☠ ☠

What is the problem? Two of the three happy individuals made it into my sample, no frowny people did, and three stoic people did. Thus, the happy group from the population is overrepresented and the frowny group from the population is unrepresented and the sample is not representative. What can I do? Select more people for the sample. Assume that I select an additional five individuals, in addition to the original five, for a sample size of *n* = 10:

Original Five ***Additional Five***

☟ ☟ ☠ ☠ ☠ ☠ ☠ ☞ ☠ ☠

You can see that I now have one frowny individual; hence, that group is represented in the population. Assume I choose another ten individuals, in addition to the *n* = 10 above, to create a sample of *n* = 20:

Original Ten

☟ ☟ ☠ ☠ ☠ ☠ ☠ ☞ ☠ ☠

Additional Ten

☠ ☠ ☠ ☞ ☠ ☠ ☞ ☠ ☠ ☠

You can now see I have two frowny individuals, two happy individuals, and a lot of stoic individuals. Thus, both small groups are represented equally in the sample, and have a smaller representation that the stoics, which was the largest group in the population. In this case, with the larger number of individuals, my sample is more representative of the population (*strength in numbers*).

The basic idea is that as you increase the sample size, you capture more and more of the individual differences in the population, which actually makes your sample more and more representative of the population. With a small sample, if you get one or two outliers, the sample is less representative than if you had a larger sample. Thus, strive for larger samples; resistance is futile!

1.14 Representative of Whom?

One issue with sampling and human data collection is what population a sample actually represents. That is, even if you engage in appropriate random selection and random assignment so that a sample should be representative of a particular population; the question remains, does the sample represent the intended population, some other population, or a sub-group within the intended population?

This is especially relevant given that most psychological research is conducted at colleges and universities, using college students as research subjects. Thus, while an investigator may use appropriate sampling techniques to ensure that subjects were selected for a sample without bias, the sample itself may never be truly representative of a population outside of the college student population.

Stated differently, assume the intended population for an investigator is "people," that is, normal, run of the mill people. If the investigator's sample consists of only college students, then one may question how well this sample represents "people." This is because college students have characteristics that may not be reflective of everyone in the human race. College students are generally of higher intelligence (especially at academically rigorous schools such as MIT and Williams College), college students often have a slight socioeconomic advantage (especially at expensive, prestigious schools such as Yale and Harvard), college students are of the same age group (18 – 23), etc. Thus, a sample of college students may not be representative of "all people," rather, a sample of college students may be representative of, well, other college students.

Even then, because the general makeup of students differs across colleges and universities, a sample of college students at a particular school may be representative of college students only at that particular school. Specifically, students attending The University of Albany (SUNY-Albany) are going to differ from students attending Boston College, who will differ from students attending Yale and MIT. Thus, a particular sample of college students may not represent "all college students," but may represent "college students from school 'X'."

Even more reductionist, the sample of college students at a particular school may represent only those students enrolled in Introduction to Psychology, or only those students who sign up for research studies. This exercise can go on and on. Thus, it becomes critical that interpretations of data collected within a population not extend beyond the actual population that is being represented by a sample.

1.15 Research Designs and Statistics

Research designs differ along a number of dimensions, but one of the main differences between designs is whether independent variables are used. All designs should make use of random-sampling, where each subject in a population has an equal chance of being selected, unless there is a reason for non-random sampling. Importantly, the choice of research design depends mainly on the research question.

Correlational Designs are used when the research question is asking whether there is a statistical association between two quantitative dependent variables. Correlational designs do not include independent variables, only dependent variables. Because no independent variables are included, you cannot make cause-and-effect relationships. All that you are doing in a correlational design is taking the scores from two or more dependent variables and seeing if they are *statistically associated*. For example: '*Is there a statistical association between SAT scores and freshman-year GPA*?' is a research question that could be answered with a correlational design. You could randomly select a sample of freshmen students

who have taken the SAT. You would obtain their SAT scores and their GPA at end of freshmen year. Through one of several statistical procedures, you could see if the SAT scores are related to the GPAs; that is, did students who scores higher on the SATs tend receive higher GPAs?

Experimental designs are used when the research question is whether there is a causal-link between two variables and you want to examine the direct influence of an independent variable on a dependent variable. Specifically, if you want to say that changes in one variable are *causing* changes in another variable, you are going to use an experimental design. In an experimental design, a researcher directly manipulates one or more independent variables and measures at least one dependent variable. The aim is to see if the manipulation of the independent variable is associated with changes in the dependent variable. If so, then the researcher may be able to conclude that the changes to the independent variable *caused* the changes to the dependent variable. An independent variable can be manipulated in one of two ways:

In **between-subjects,** each subject is randomly assigned to only one level of an independent variable. For example, a researcher may be interested in whether a new antidepressant drug is effective in alleviating depression. The researcher could randomly assign subjects to take either the drug, or an inert placebo; thus, the researcher has manipulated whether a subject takes the drug or a placebo (independent variable). The researcher would measure depression and compare depression levels between these two groups. In **within-subjects** the same group of subjects is tested under each level of the independent variable; hence, each subject is tested several times. For example, a researcher may be interested in whether background noise affects test taking. The researcher could randomly select a sample of students and have them take some standardized test in a *noisy condition*, and then test these same subjects in a *quite condition*. Thus, noise level has been manipulated (independent variable) and each subject is expose to each level of this independent variable. The researcher would then compare test scores between conditions.

Quasi experimental designs are used when the research question is whether there is a difference in performance between two or more preexisting groups of subjects. The groups differ along an independent variable, but the independent variable is not manipulated and subjects already belong to a group. For example, *sex* (male versus female) is an independent variable that cannot be manipulated, because a subject naturally falls into one of the two levels. Because the independent variable is not being manipulated, a researcher cannot establish cause-effect relationships; rather, the researcher can only establish group differences. For example, a researcher may want to know whether males differ in their spatial reasoning skills from females. The researcher would randomly select a group of males and a group of females, and then test each male subject's spatial reasoning skills and test each female subject's spatial reasoning skills. The performance between these two groups would then be compared.

1.16 Computers and Statistics

There are a variety of computer programs that allows you perform statistical analyses and procedures on a set of data (e.g., SPSS/PASW, Minitab, SAS, R, MATLAB, and Microsoft Excel). Indeed, all of the procedures that we will cover in this course can be performed using software. So why the heck are you going to be learning to calculate various statistical parameters and perform various statistical procedures by hand? Why not just do it all by computer?

The simple explanation is you need a foundational knowledge of what the various statistics you will encounter represent, what they measure, what come from, and how they are calculated. I can easily show you how to perform something called a 'between-subjects factorial analysis of variance' using several of the software programs mentioned above; however, if you do not know anything about this procedure, and what it means to 'analyze variance', then you will have no idea how to interpret the results of such an analysis done by way of one of these programs.

I will be teaching you the nuts and bolts of various procedures by having you perform analyses by hand and by calculating parameters by hand. This will allow you to see where things like 'variance' come from, what the various measures of 'variance' reflect, and when each is to be used. For certain topics, we'll then

go over how to have a software program calculate these parameters for us. Thus, you must understand the conceptual framework first, before you can hope to interpret computer-generated output.

CH 1 Homework Questions

1. *Identify each of the following as a variable or a constant. Explain the reasons for your choices*.

a. The number of days in a week.
b. People's attitudes toward abortion.
c. The country of birth of presidents of the United States.
d. The value of a number divided by itself.
e. The total number of runs scored in baseball games.
f. The number of days in each of the twelve months in a year.

2. *Identify each of the following as a variable or a constant. Explain the reasons for your choices*.

a. An individual's attitude toward abortion at a specific point in time.
b. The number of days in February.
c. Peoples' opinion of the death penalty.
d. Grade Point Average (GPA).
e. The number of hairs on someone's head.
f. The time it takes to complete a homework assignment.
g. A student's semester GPA at the end of the semester.
h. The number 12

3. *Identify each of the following as a qualitative or a quantitative variable:*

a. age
b. religion
c. yearly income
d. weight
e. gender
f. eye color
g. college major
h. political party
i. temperature

4. *Identify each of the following as a qualitative or a quantitative variable.*

a. a person's name
b. goals scored by a hockey team
c. length of a rope
d. shoe color
e. movie titles
f. duration of a movie
g. number of licks to get to the center of a Tootise Roll pop
h. University attended
i. brain activity as measured via EEG
j. camera price

5. *Define each of the following, in <u>your</u> <u>own</u> words.*

a. Population
b. Sample

c. Null hypothesis
d. Alternate hypothesis

6. *Define each of the following, in your own words.*

a. Independent variable
b. Dependent variable
c. Extraneous variable
d. Confounding variable
e. Control variable

7. How is a sample used to make conclusions about the population?

8. What is the main difference between an experimental design and a correlational design?

9. *Answer each of the following, in your own words.*
a. What are descriptive statistics used for?
b. What are inferential statistics used for?
10. *Answer each of the following, in your own words.*
a. What is the difference between a statistic and a parameter?
b. What types of symbols are, typically, used for statistics and parameters?

11. Distinguish between a dependent variable and an independent variable.

12. Distinguish between a between-subjects variable and a within-subjects variable.

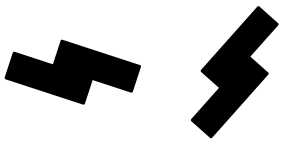
For Exercises 13 – 16, identify the independent variable and the dependent variable in the scenario, and indicate whether each variable is quantitative or qualitative.

13. An antidepressant drug has been shown to have positive results on alleviating depression. To examine the drug's effectiveness on reducing depression, Dr. Leary administers different dosages of the drug (0-mg, 10-mg, 20-mg, 30-mg, 40-mg, and 50-mg) to six different groups, and then measures the amount of brain activity in each individual; where brain activity is a measure of depression.

14. In a classic study on aggression, Eron (1963) studied the relationship between the amount of violent television shows watched by young children and the amount of aggressive behavior they showed toward peers. Eron questioned the parents of over 800 children about their child's television viewing habits; and Eron created a four-point scale to measure preference for violent television. Aggressive behavior was measured by collecting ratings of each child by two or more children who knew the child in the study. Ratings ranged from 0 - 32, with higher scores indicating more aggressive behavior.

15. Rosenthal and Fode (1963) examined the effect of experimenter bias on the ability of rats to learn to navigate a maze. Subjects were told that they were going to teach a rat to learn to navigate a maze. One group of subjects was told that their rats were genetically smart, "maze bright" rats that should show learning during the first day and performance should increase. A second group of subjects was told their rats were genetically dumb, "maze-dull" rats that should show very little evidence of maze learning. Subjects trained their rats to enter one of two sections of a maze, by rewarding the rat when they entered the correct section. Subjects measured the number of times the rat entered the correct section of the maze each day.

16. Shepard and Metzler (1971) studied the ability of humans to 'mentally rotate' the image of an object. Subjects were shown two pictures, like those to the right, and then had to decide whether the two objects were the same shape, or were mirror images of the same shape. The angle of rotation separating the two objects varied from 0° to 180° in 15° increments. Subjects were assessed on how quickly they could correctly respond whether the objects were the same, or were mirror images.



17. What is the main purpose of random selection?

18. What is the main purpose of random assignment? When is it used?

19. Give two examples of a population and a sample of that population. For each of your examples, be sure that the sample is related to the population.

20. You have a population that includes 1000 individuals and each member of this population is assigned a number from 1 to 1000. Using the website http://www.randomizer.org/, select a random sample of $n = 20$ individuals and write the numbers on your answer sheet.

21. Repeat the random selection exercise from #20 and write the numbers of the selected individuals on your answer sheet. How many of the same individuals were selected to participate in both samples?

22. If the population in #20 and #21 included only 100 individuals, would you be more or less likely to randomly select the same person twice?

23. *Use the following scenario to answer a – f below*: Dr. Evil has given up evil to start a potato chip company that will distribute potato chips to undergraduates in the United States. Dr. Evil wants to know what potato chip flavor undergraduates like best. He asks 100 undergraduates to taste each of four potato chip flavors (1) BBQ, (2) Plain, (3) Ranch, and (4) Super-Evil. Each student gives a rating on a scale of 1 to 10 on how much they like each flavor, with 10 indicating they really like the flavor. The average ratings for each of the four chip flavors are 4.5 for BBQ, 6.7 for Plain, 2.5 for Ranch and 9.5 for Super-Evil.

a. The population in this scenario is...
b. The sample in this scenario is...
c. “Fred” rates the Super-Evil Flavored chips a 10. This rating is a...
d. The average ratings in the scenario is(are) a...

24. *Use the following scenario to answer a – e below*: Dr. Logan is interested in the relationship between playing violent video games and aggression in adolescent males. He randomly selects 1000 adolescent males from around New York city and administers a survey for the types of video games that each adolescent male likes to play. He then obtains measures of each adolescent male's aggression by asking each adolescent male's teacher to provide a "toughness rating" on a scale from 1 to 10. The average toughness rating for these adolescent males is 4.56.

a. The population is...
b. The sample is the...
c. Say that "Timmy," one of the adolescent males, scores a 7 on the toughness rating. This value of 7 is a...
d. The average toughness rating is a...
e. Is the types of video games that each adolescent male likes to play a qualitative or quantitative variable?

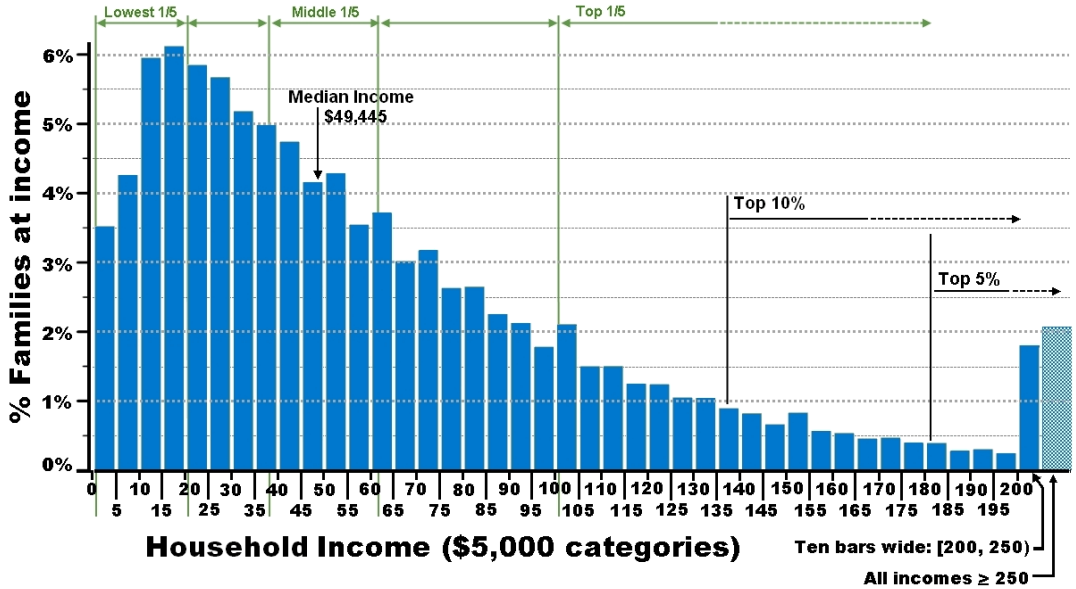
25. *Use the following scenario to answer a – g below:* Dr. Smith wants to determine whether playing music during class causes statistics students to pay more attention. He has three stats classes with 40 students/class. For one semester he teaches his morning class while playing Snoop Dogg, in his afternoon class he plays Dio and in his night class he plays no music (a control). During each class meeting, Dr. Smith counts the number of students that fall asleep. He finds an average of 6 students per meeting fall asleep in his morning class, 2 students per meeting fall asleep in his afternoon class and 5 students per meeting fall asleep in the night class.

a. The population in this scenario is...
b. The sample in this scenario is...
c. In one of Dr. Smith’s night classes, he counts 14 students that fall asleep. This number is...

d. The average of 6 students that fall asleep in the morning class is...
e. The use of different music in each class is a...
f. Counting the number of students that fall asleep uses what kind of measurement scale?
g. Before Dr. Smith can draw any conclusions about the effect of music on statistics student's paying attention in class, Dr. Smith will need to use...

26. The town of Petoria (population = 75000) is holding a local election for mayor between candidates Bill Hicks and Peter Swanson. The local newspaper conducts a survey by asking readers to indicate their preference for Bill Hicks or Peter Swanson, by sending in a ballot provided in the paper with their preference indicated. One week later the newspaper reported it received 2500 ballots, of which 2000 favored Bill Hicks. The next day, newspaper printed a story claiming that Bill Hicks would win a landslide in the upcoming election. Based on this information, is the newspaper's sample representative of Petoria? Why or why not? What implication does this have for the newspaper's prediction about the election?

Chapter 2: Frequency Distributions and Graphing



Data source: http://www.census.gov/hhes/www/cpstables/032011/hhinc/new06_000.htm

This chapter addresses some of the basic uses of statistics, that is, organizing and summarizing data through frequency distributions and graphing. Before discussing these topics, some other issues must first be addressed that deal with types of data, which has implications for how data are organized and graphed.

2.1 Scales of Measurement

Data comes in many forms, but most forms of data can be classified as belonging to one of four measurement scales. I start with the least complex scale and move to the more complex scales where each increasingly complex scale has the qualities of the scales below plus something additional.

Nominal scale data are qualitative and categorical, with the measurements being associated with distinct groups or events; hence, nominal measurements have no numerical values and there is no order to the levels on a nominal scale. For example, the table below lists the two divisions in each the Eastern Conference and the Western Conference that make up the of the National Hockey League (NHL), along with the teams within each division. Each team's *division* is measures on a nominal scale, as each of the four divisions can be measured, but have no numeric value.

Eastern Conference		**Western Conference**	
Atlantic Division	**Metropolitan Division**	**Central Division**	**Pacific Division**
Boston Bruins	Pittsburgh Penguins	Dallas Stars	Edmonton Oilers
Montreal Canadiens	Philadelphia Flyers	Chicago Blackhawks	Calgary Flames
Detroit Red Wings	Carolina Hurricanes	Minnesota Timberwolves	Los Angeles Kings
Buffalo Sabres	Washington Capitals	Winnipeg Jets	Arizona Coyotes
Ottawa Senators	New York Islanders	Nashville Predators	San Jose Sharks
Toronto Maple Leaves	New Jersey Devils	St. Louis Blues	Anaheim Ducks
Florida Panthers	New York Rangers	Colorado Avalanche	Vancouver Canucks
Tampa Bay Lightning	Columbus Blue Jackets		

An **ordinal scale** is like a nominal scale because the measurements are generally categorical, however measurements on an ordinal scale have structure and order that allows one to determine whether each

measurement is better/worse or greater/less than other measurements. The order of entries is based on a relative measure, but the measure itself is meaningless, because there is no value associated with each entry. For example, the table below lists the same data as in the nominal scale table above; however, I listed the teams within each division in the order in which they finished the 2014-2015 NHL season:

Eastern Conference		**Western Conference**	
Atlantic Division	**Metropolitan Division**	**Central Division**	**Pacific Division**
1st: Montreal Canadiens	1st: New York Rangers	1st: St. Louis Blues	1st: Anaheim Ducks
2nd: Tampa Bay Lightning	2nd: Washington Capitals	2nd: Nashville Predators	2nd: Vancouver Canucks
3rd: Detroit Red Wings	3rd: New York Islanders	3rd: Chicago Blackhawks	3rd: Calgary Flames
4th: Ottawa Senators	4th: Pittsburgh Penguins	4th: Minnesota Timberwolves	4th: Los Angeles Kings
5th: Boston Bruins	5th: Columbus Blue Jackets	5th: Winnipeg Jets	5th: San Jose Sharks
6th: Florida Panthers	6th: Philadelphia Flyers	6th: Dallas Stars	6th: Edmonton Oilers
7th: Toronto Maple Leaves	7th: New Jersey Devils	7th: Colorado Avalanche	7th: Arizona Coyotes
8th: Buffalo Sabres	8th: Carolina Hurricanes		

You can see the order of the teams within each division is based on their place of finish; however, the ordinal scale does not tell us how much better or worse any one team did than other teams. An ordinal scale provides information about how one measurement (i.e., team) performed *relative* to others.

An **interval scale** is similar to an ordinal scale because the data are ordered, but unlike an ordinal scale, measurements on interval scales are quantitative, which can be used to determine numeric differences between measured values. Importantly, the interval between adjacent values is equal. For example, the difference between the values 2 and 3 is 1 on an interval scale, and the difference between the values of 4 and 5 on that same scale is also 1.

Interval scales do not have a **true zero point**, that is, a value at which the measured variable has no magnitude. A true value of zero tells you there is nothing of that variable present when it is being measured. Stated differently, an an interval scale a value of 0 can be obtained, but that value of 0 is actually meaningless, because the 0 still means something there. Hence, interval scales often have negative values as possible measurements, and those negative values mean something. Because there is no true zero point, you cannot form ratios and cannot tell how many times greater-than/less-than one value is from another. In order to form ratios you need to have a true zero point, because there can be no such thing as 'negative ratios'. One example of something that is measured on an interval scale is measuring temperature in degrees fahrenheit, because you can have 1°F, 0°F, -1°F, -2°F... etc.

The **ratio scale** is identical to an interval scale except the variable being measures is one that has a true zero point, because negative values are not possible and are meaningless on ratio scales. With a true zero point a value of 0 indicates the absence of any magnitude. I should note that the true zero point does not have to show up in the data and is unlikely to ever show up. For example, any measurement of time (days, seconds, hours, milliseconds) is on a ratio scale, because you cannot have negative time.

To which scale of measurement do psychological constructs belong? Although the actual measures of psychological constructs may be on ratio scales (e.g., reaction times are ratio, accuracies are ratio); most constructs are actually measured on ordinal scales and nominal scales. For example, many IQ tests are standardized to have an average score of 100 with a theoretical range of 0-175. This appears to be a ratio scale, because you certainly cannot have 'negative intelligence', but, take two IQ scores, 50 and 100. Is the individual with the score of 100 double the intelligence of the individual with the IQ score of 50? Absolutely not. All that can be said is that the individual scored 50 points higher on the IQ test; hence, IQ scores seem to be on an interval scale, but in reality we would only conclude the individual with the IQ of 100 is more intelligent; hnce, the IQs would fall along an ordinal scale.

Also, recall from Chapter 1 a Likert scale (see below) appears to be on an interval scale because of the numerical values, but is actually on an ordinal scale, because people are choosing between different categories placed relative to each other. Thus, there can be a difference between the scale in which a variable can be interpreted, and the actual scale by which a variable is measured.

Overall, what is your general political attitude?

-4	-3	-2	-1	0	+1	+2	+3	+4
very liberal	moderately liberal	somewhat liberal	slightly liberal	totally neutral	slightly conservative	somewhat conservative	moderately conservative	very conservative

2.2 Continuous vs. Discrete Data

For quantitative data that has been measured on an interval or a ratio scale, the data can be continuous or discrete. **Continuous data** have a theoretically infinite number of possible values between any two measurable points, that is, data are assumed to continue infinitely between any two measurable values, so a measurement that contains a decimal or is a fraction of a whole number makes sense. For example, between 1 second and 2 seconds there can be 1.02 seconds, 1.156 seconds, 1.5432678 seconds, etc. Each of those values makes sense because each value can exit.

In contrast, with ***discrete data*** there are a finite or fixed number of values that can exist between any two measurable points; hence, decimals and fractions of measurable numbers do not make sense. For example, one row of desks in a classroom may contain 2 students and a second row may contain 4 students. The only possible measurement for people between 2 students and 4 students is 3 students, because you cannot have a measurement of say 2.5 people or 3.75 people.This is not to say that you cannot have an average of 4.5 people sitting in each row of a classroom! With discrete data, individual measurements cannot be in fraction or decimal form, but the average is a different story. Indeed, an average of 4.5 people sitting in each row simply indicates that there are, on average, 4 or 5 people sitting in each row.

2.3 Real vs. Apparent Limits of Continuous Data

With continuous data any measurement is only as good as the device making the measurement and as good as the person who is taking the measurement. Thus, when we say that someone's reaction time to push a button was 1.2 seconds, we do not mean exactly 1.2 seconds; rather, we mean approximately 1.2 seconds. Similarly, when I say that it took me 5 years to complete graduate school to receive my Ph.D., I do not mean 5 years exactly; I mean 5 years approximately.

If you are rounding numbers, you are taking away some of the accuracy of the measurements. For example, if you are rounding to the tenths place and you measure someone's actual reaction time to be 1.56, you would round that to 1.6. Thus, you have introduced inaccuracy. Thus, when we say that we have measured a person's reaction time to be 1.6 seconds, we may actually mean that the reaction time was 1.56 seconds or 1.61 seconds, or even 1.599999 seconds.

To account for this approximation, we place **real limits** around values and consider any value that lies within these real limits to be associated with the listed value. Thus, real limits are like a boundary around a reported value that includes similar values that we equate with that reported value. Any value within that boundary will be listed as the number around which the boundary is based.

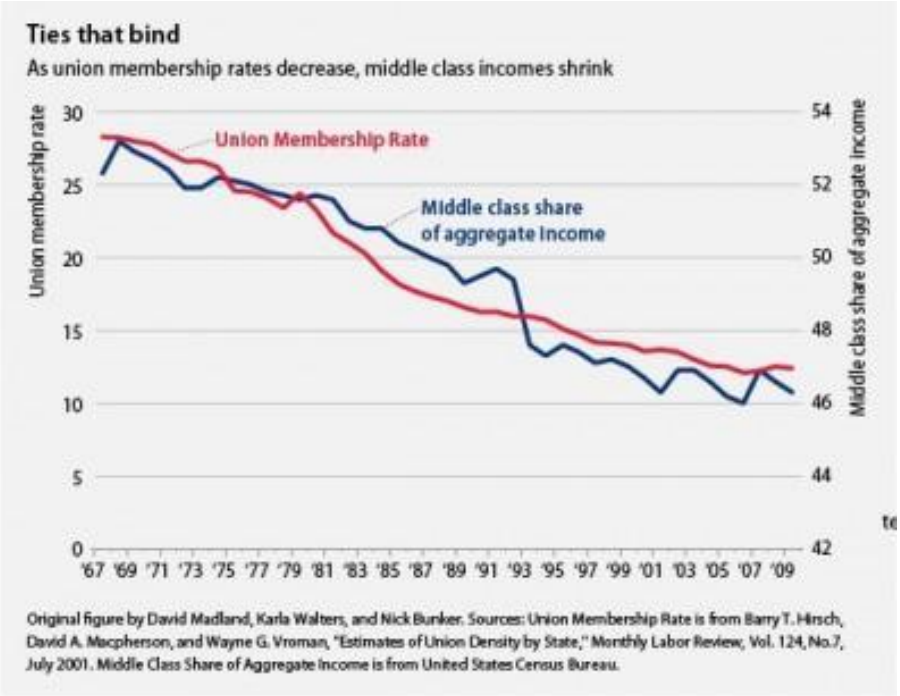
Real limits of any number are defined as the values that fall one-half unit of measure above that number and one-half unit below that number. For example, the real limits of 19 are 18.5 and 19.5. The smaller value is the **lower real limit** (*LRL*) and the larger value is the **upper real limit** (*URL*). Similarly, the real limits around 1.2 are 1.25 and 1.15. To find the real limits of any number, follow these steps:

Step	**What to Do**	**9**	**9.0**	**9.00**

(1)	Determine the lowest unit of measure	Ones	Tenths	Hundredths
(2)	Take one unit of that lowest unit of measure	1	0.1	0.01
(3)	Divide that one unit in half	1/2 = 0.5	0.1/2 = 0.05	0.001/2 = 0.005
(4)	Add the value in step 3 to the original number to get the upper real limit	URL = 9 + 0.5 URL = 9.5	URL = 9.0 + 0.05 URL = 9.05	URL = 9.00 + 0.005 URL = 9.005
(5)	Subtract the value in step 3 from the original number to get the lower real limit	LRL = 9 - 0.5 LRL = 8.5	LRL = 9.0 - 0.05 LRL = 8.95	LRL = 9.00 - 0.005 LRL = 8.995

2.4 Frequency Distributions for Quantitative Variables

When data are collected they are unorganized, and one of the basic uses of statistics is to organize raw data. The graph below is an example of raw data can be consolidated into a simple figure to express information. This chapter is concerned with taking unorganized raw pieces of information and organizing them into something meaningful.



A **frequency distribution** is data that has been organized into a table or a graph. In table form each value or level of a variable is listed with the number of times that value or level appeared in the data. As a graph, each value of level of the variable is presented as a function of the number of times that value or level occurred in the data. Frequency distributions simply consolidate a data set into something smaller and manageable. Generally, **frequency distribution table** (or simply **frequency table**) refers to a frequency distribution in table form; whereas **frequency distribution** refers to one in graphed form.

Let's say we ask *n* = 50 students the following question and have each student rate their response to the question on a scale from 1 to 11: *Overall, what is your general political attitude*? The rating scale is below:

1	2	3	4	5	6	7	8	9	10	11
extremely liberal	very liberal	moderately liberal	somewhat liberal	slightly liberal	totally neutral	slightly conservative	somewhat conservative	moderately conservative	very conservative	extremely conservative

I obtain the following scores:

5	6	8	9	6	5	7	6	5	6
5	4	5	5	6	7	6	8	2	3
4	6	8	11	6	7	5	3	2	6
6	8	7	9	5	6	11	3	2	5
6	7	9	5	6	7	8	9	8	3

In present form these data are unorganized and it is impossible to draw conclusions or make inferences about political attitudes. Is there anything wrong with the data being unorganized? No, the data are fine, but trying to understand the data is difficult when presented this way. The simplest method for organizing data is to create a frequency distribution table, where each value in the data set is listed with its **frequency** of occurrence, that is, the number of times each value was recorded in the data. The most important thing to remember is to make the table (or graph) easy to understand.

To construct a frequency table, first list each value from high to low in one column by placing the highest value at the top of this column and working your way down to the lowest value. The name at the top of this column is usually the name of the variable ("Political Attitude"). You can skip missing values in the range or not skip them; this is really a judgment call. For example, a rating of 10 was not recorded in the data above, but because 10 is in the range of possible values between the highest obtained score (11) and lowest score (2), you might want to include the value of. One reason for including values not actually obtained, but within the range of possible values, is to prevent people from wondering about missing data. In the frequency table below I included the values 1 and 10 which were not recorded in the data above, but are possible values. Next, count the number of times each value occurred in the data, this is the value's **frequency**, and list the frequency of each value in a new column labeled 'f' (see table below). It is good to list the **total frequency** (n) at the bottom of this column.

Political Attitude	**f**
11	2
10	0
9	4
8	6
7	6
6	13
5	10
4	2
3	4
2	3
1	0
	n = 50

Except for some additional information to be added, that's it; this is a simple frequency distribution table. You can now easily see most students rated their overall political attitude as 5 (slightly liberal) or as 6 (totally neutral), and you could conclude the population from which these students came is not overly liberal or overly conservative, but is more centrist. Without the data being organized in this manner, such an inference is much more difficult to make.

There is a limitation to listing **absolute frequencies** as these don't tell you anything about the rest of the distribution, that is, absolute frequencies don't tell you how frequent a value occurred relative to the rest of the distribution. For example, if I told you that four students reported a political attitude of 9 (moderately conservative), but gave no other information, what does this say about a value of 9 with respect to the distributio? Knowing that four students reported having a moderately conservative political attitude does not tell you anything about whether this is a high frequency or a low frequency relative to the total frequency. A frequency of four would be high if there were only 10 students, but would be low if there were 10,000 students. If you don't know the total number of students (n), you cannot determine whether four students is a high or a low frequency.

We need a standard measure that tells us of the frequency of a value relative to the distribution such that higher values indicate a higher impact and lower values indicate a lesser impact. This is **relative frequency** (**rf**) is the **proportion** of times each value occurred in the data. The relative frequency of a value is calculated by dividing the absolute frequency of a value by the total frequency (n) and listing that resulting value in a column to the right of the frequency column:

Political Attitude	f	rf
11	2	2/50 = .04
10	0	0/50 = .00
9	4	4/50 = .08
8	6	6/50 = .12
7	6	6/50 = .12
6	13	13/50 = .26
5	10	10/50 = .20
4	2	2/50 = .04
3	4	4/50 = .08
2	3	3/50 = .06
1	0	0/50 = .00
	n = 50	

You do not include the quotients in the relative frequency column; only include the proportions. What's nice about using relative frequency is that proportions range from 0 to 1 so they provide a standard measure of impact that a value has relative to the distribution. That is, how much impact does a value have on a distribution? The higher the relative frequency the greater impact that value has on the distribution. A similar measure that accomplishes the same goal is the **relative percentage** (**%**) of a value. The relative percentage is the relative frequency multiplied by 100.

Political Attitude	f	rf	%
11	2	.04	.04 x 100 = 4%
10	0	.00	.00 x 100 = 0%
9	4	.08	.08 x 100 = 8%
8	6	.12	.12 x 100 = 12%
7	6	.12	.12 x 100 = 12%
6	13	.26	.26 x 100 = 26%
5	10	.20	.20 x 100 = 20%
4	2	.04	.04 x 100 = 4%
3	4	.08	.08 x 100 = 8%
2	3	.06	.06 x 100 = 6%
1	0	.00	.00 x 100 = 0%
	n = 50		

Frequency distributions are also useful for determining the 'place' or 'rank' of a value relative to other values. Specifically, frequency distributions can be used to determine the number, proportion, or percentage of scores greater than, less than, greater than or equal to, or less than or equal to a value. To do this we must add several other pieces of information to the frequency distribution.

The **cumulative frequency** (**cf**) is the number of scores less than or equal to a value, so it is the frequency of a value added to the frequencies of all values less than it. For example, take the political attitude rating of 4 (somewhat liberal). The cumulative frequency of a rating of 4 is its frequency (f = 2) added to the frequency for the ratings of 3 (f = 4), 2 (f = 3), and 1 (f = 0). This results in a cumulative frequency of 9 for the rating of 4. This indicates there are 9 scores less than or equal to 4. Stated differently, 9 students have a political attitude that is 'extremely liberal' to 'somewhat liberal'.

To determine the cumulative frequency for each value in a frequency distribution always start with the smallest value and work through larger values. First, determine the cumulative frequency of the value 1 by adding its frequency (0) to the frequency of all values less than 1. Next find the cumulative frequency

for the next largest value (2) by adding its frequency (3) to the frequency of all values less than 2. Continue this for each increasing value until you have filled the cf column:

Political Attitude	f	rf	%	cf
11	2	.04	4%	2 + 48 = 50
10	0	.00	0%	0 + 48 = 48
9	4	.08	8%	4 + 44 = 48
8	6	.12	12%	6 + 38 = 44
7	6	.12	12%	6 + 32 = 38
6	13	.26	26%	13 + 19 = 32
5	10	.20	20%	1 + 9 = 19
4	2	.04	4%	2 + 7 = 9
3	4	.08	8%	4 + 3 = 7
2	3	.06	6%	3 + 0 = 3
1	0	.00	0%	0 + 0 = 0
	n = 50			

The cumulative frequency tells you nothing about what proportion of scores are less than or equal to a value. To determine this you need the **cumulative relative frequency** (**crf**) of a value, which is obtained by dividing the cumulative frequency by the total frequency (n). In the table below, each value in the crf column is the proportion of scores in the distribution that are less than or equal to the value in the Political Attitude column. You can also convert the cumulative relative frequency values into **cumulative percentages** (**c%**) by multiplying each crf by 100 and adding the percent sign. In the table below, I have added the cumulative relative frequencies and the cumulative percentages. Again, you do not have to include the mathematical operations in the crf and c% columns:

Political Attitude	f	rf	%	cf	crf	c%
11	2	.04	4%	50	50/50 = 1.00	1.00 x 100 = 100%
10	0	.00	0%	48	48/50 = .96	.96 x 100 = 96%
9	4	.08	8%	48	48/50 = .96	.96 x 100 = 96%
8	6	.12	12%	44	44/50 = .88	.88 x 100 = 88%
7	6	.12	12%	38	38/50 = .76	.76 x 100 = 76%
6	13	.26	26%	32	32/50 = .64	.64 x 100 = 64%
5	10	.20	20%	19	19/50 = .38	.38 x 100 = 38%
4	2	.04	4%	9	9/50 = .18	.18 x 100 = 18%
3	4	.08	8%	7	7/50 = .14	.14 x 100 = 14%
2	3	.06	6%	3	3/50 = .06	.06 x 100 = 6%
1	0	.00	0%	0	0/50 = .00	.00 x 100 = 0%
	n = 50					

The table above is a complete frequency distribution table that organizes raw data, provides information about the proportion each value contributes to a distribution, and provides information about the numbers and proportions of scores less than or equal to a value. Such a distribution provides a means for someone to assess the importance of the data and to discuss the data and possible make some inferences. Now that this frequency distribution table has been established for the political attitude data from earlier, we can start to answer some general questions about this data. The examples that I list below, are just several of questions that could be answered from a frequency distribution table.

Question: How *many* students rated their political attitude as a 5 (slightly liberal) or less? *Answer*: 19--This comes for the cumulative frequency of 5.

Question: How *many* students rather their political attitude as being less than a 5 (slightly liberal)? *Answer*: 9--This comes from the cumulative frequency of 4, which is the immediately less than 5.

Question: What *proportion* of students rated their political attitude as a 9 (moderately conservative)? *Answer*: .08--This comes from the relative frequency of 9.

Question: What *proportion* of students rated their political attitude as an 8 (somewhat conservative) or less? Answer: .88--This comes from the cumulative relative frequency of 8.

Question: How *many* students rated their political attitude as greater than 7 (slightly conservative)? *Answer*: 12--This can be found by adding the absolute frequencies of the values greater than 7 (i.e., 8, 9, 10, and 11); or, by subtracting the cumulative frequency of 7 (crf = 38) from the total frequency (n_T = 50).

2.5 Frequency Distributions for Qualitative Variables

The frequency distribution table created in Section 2.4 was based on quantitative data; the numerical ratings the students gave for political attitudes. However, it may be that a variable is qualitative and the data on a nominal scale. For example, a professor may want to know want to know the numbers of college majors within a class or the numbers of freshmen, sophomores, juniors, and seniors within a class. How would a frequency table for qualitative/nominal data be set up?

Say I want to know the numbers of each college major in a particular class with 30 students. I ask each student his/her primary major and list each major with the number of students in that major. In this case, each 'value' for the variable College Class is qualitative and the data are on a nominal scale. When creating a frequency distribution for qualitative data, most steps are identical to those when creating a frequency distribution with quantitative data:

College Class	f
Psychology	15
Neuroscience	5
English	2
Biology	4
History	4
	n = 30

The only difference between frequency distributions for quantitative data and for qualitative data is you do not include cumulative frequency, cumulative relative frequency, and cumulative percentages for qualitative data. Frequency distributions for qualitative data can include the relative frequencies and the relative percentages for each entry. I have listed these in the table below:

College Class	f	rf	%
Psychology	15	.500	50%
Neuroscience	5	.167	16.7%
English	2	.067	.6.7%
Biology	4	.133	13.3%
History	4	.133	13.3%
	n = 30		

2.6 Outliers in Frequency Distributions

One of the uses of frequency distributions is identification of **outliers**. An **outlier** is any score with a value very different from the other data value and usually has a low frequency. For example, say that I measure the age in years of each person in a study. The frequency distribution is below, where it can be seen that most of the 80 people ranged from 18 to 21 years old, which you would expect if this study was conducted on a college campus. However, there is one individual who is 45 years old; an age over twice as old as the next oldest people in the study. This individual's age (and the individual) would be an outlier.

Age	f
45	1
21	15
20	20
19	25
18	19

Outliers can be problematic, because if the outlier is a person, their individual characteristics and performance in tasks might not be typical of the rest of the sample and, hence, may be non-representative of the population. As such, outliers are often eliminated from statistical data analysis.

Outliers can also be identified when data come from a qualitative variable such as college major. For example, say I conduct a survey on students' opinions of campus events and obtain the following distribution of majors in my sample:

College Class	f
Psychology	50
Philosophy	2
English	45
Biology	90
History	45

In this case, the two philosophy majors would be considered outliers and their data may be removed, as their responses may not be in line with the majority of the sample.

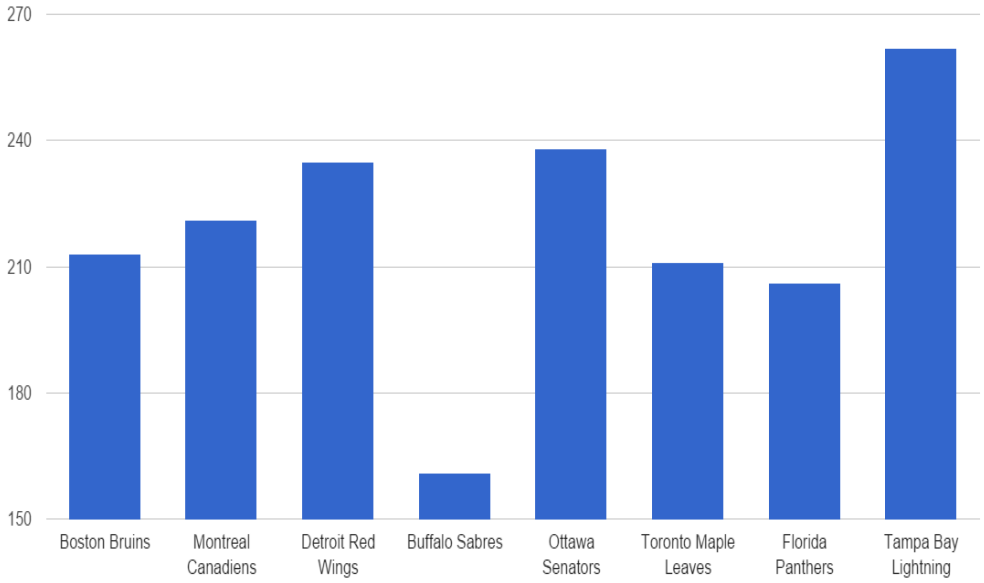
2.6 "A Picture is worth a Thousand Words"

I'm sure you've heard this saying, but did you ever think about what it means? Basically, a picture can take the place of the roughly 1000 words needed to describe the picture. Not that describing something is bad, it's just that understanding something through seeing is often easier than having it explained. Meaning can often get "lost in translation."

The same is true in statistics: A graph of data is probably worth 1000 words, so trends and relationships in data are usually explained through graphs. Some of this section will likely be a refresher, but there are some general rules that should be followed when creating graphs. The main rule is to present data in the simplest and most meaningful way possible, but what does this mean?

1. Don't make over-complicated graphs
2. Include only what is necessary
3. Include enough information to make your graph easily interpretable

You don't want people thinking too much about what your graph means; rather, you use a graph to help people understand data. Too much 'stuff' in a graph makes it confusing, but not enough will make it unclear. For example, say we want to graph of the number of goals scored by each team in the Atlantic Division of the National Hockey League for the 2014-2015 season. There are eight teams in that division and the numbers of goals each team scored are display in the graph below, where it can be clearly seen which teams scored more goals than the others.



The dependent variable is usually displayed on the vertical, **y-axis** (**ordinate**). In the example above, the dependent variable is Total Goals Scored and the values are presented on the vertical axis. The horizontal, **x-axis** (**abscissa**) can display levels of an independent variable or levels of another dependent variable. In this case, the variable Atlantic Division Team is a qualitative variable with five different levels (teams). The title on each axis should be short and be descriptive to accurately reflect the variable on that axis. The values levels on the y-axis are up to you, but you generally do not want to have too great a range of values or too small of a range. This can lead to a very misleading graph.

2.8 Graphs for Frequency Distributions of Quantitative Variables

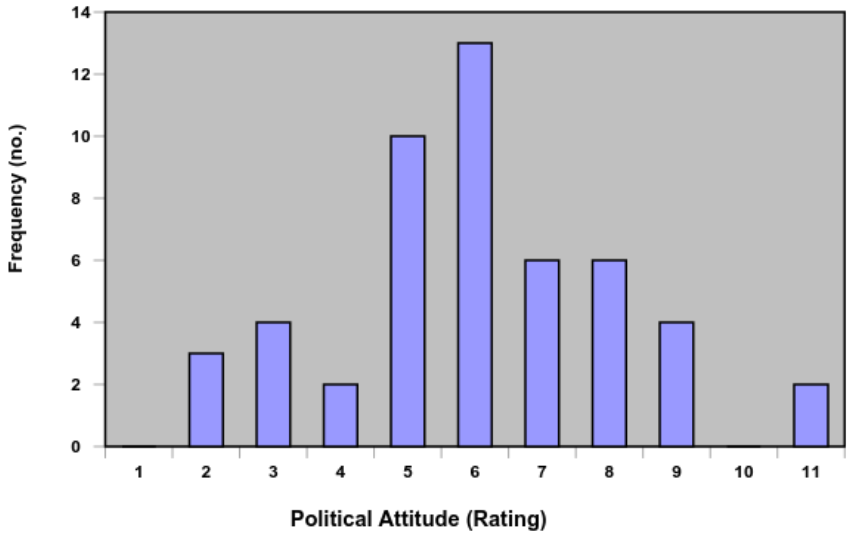
When data are measured on a quantitative variable one of three different graphs are generally used to display the data: **histograms**, **polygons**, and **line plots**. There are some rules for when each of these graphs is used, but which graph type is used is a little less important than displaying data in a clear and simple manner. For each of the following examples, I use the frequency distribution of the Political Attitude data form Section 2.4, which is reproduced below:

Political Attitude	f	rf	%	cf	crf	c%
11	2	.04	4%	50	1.00	100%
10	0	.00	0%	48	.96	96%
9	4	.08	8%	48	.96	96%
8	6	.12	12%	44	.88	88%
7	6	.12	12%	38	.76	76%
6	13	.26	26%	32	.64	64%
5	10	.20	20%	19	.38	38%
4	2	.04	4%	9	.18	18%
3	4	.08	8%	7	.14	14%
2	3	.06	6%	3	.06	6%
1	0	.00	0%	0	.00	0%

People use the terms *histogram* and *bar graph* interchangeably, but they are not the same graph. Bar graphs are used when the variable on the abscissa is qualitative, whereas histograms are used when the variable on the abscissa is quantitative. In bar graphs the bars above each category do not touch, but in histograms the bars do touch. For histogram, each bar is centered above its level on the abscissa and the

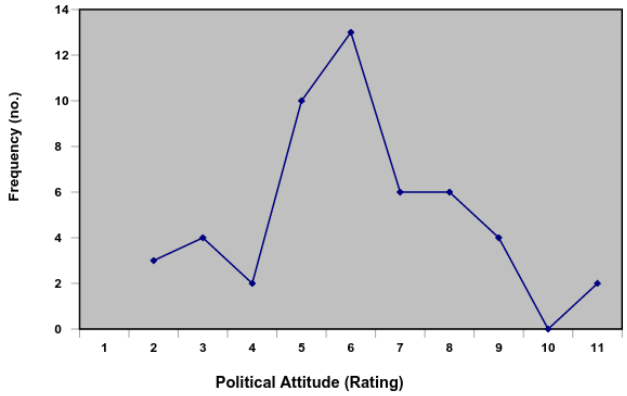
width of each bar represents real limits around each value. The height of each bar represents the magnitude of the dependent variable, which in this case is the frequency of each value.

To create a histogram set up the ordinate and the abscissa by giving each axis a proper title. Next, label the entries on each axis. Next, extend a column up to a point above each value on the abscissa that corresponds to the frequency of that value on the abscissa. Again, make sure that the bars touch.

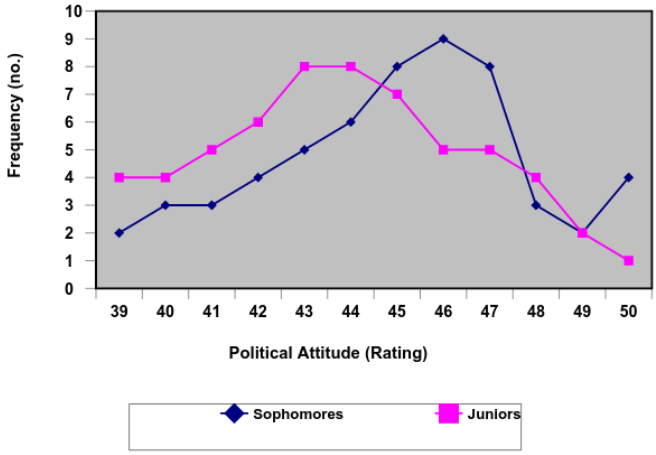


It is important to place the smallest values on the left side of the abscissa nearest the ordinate. The point where the abscissa and the ordinate meet, called the **origin**, represents the smallest values. This way, as distance increases from the origin on either axis, it indicates greater and greater values.

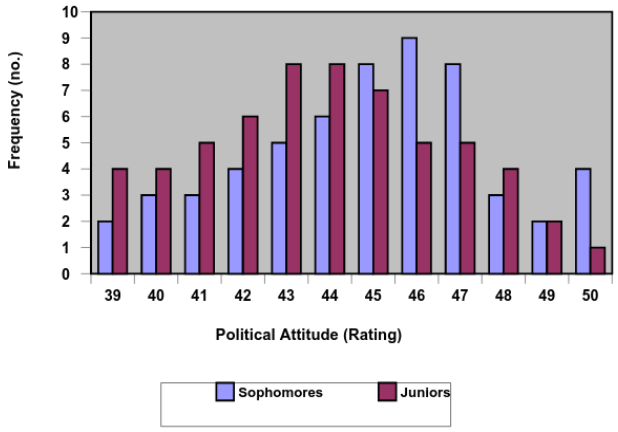
A **polygon** uses the height of lines between data points to represent the magnitude of a dependent variable, instead of columns as in bar graphs and histograms. Polygons are used when the variable on the abscissa is quantitative and is general used when the variable is continuous. Remember from Section 2.2, continuous data is where there are theoretically an infinite number of possible measurements between any two points; hence, fractional values are possible. The use of a line to connect data points in a polygon graph is supposed to represent continuity between the data points. The graph below presents the Political Orientation data as a polygon.
You'll notice that I added a score of 12 to the abscissa and gave it a frequency of zero. This is because frequency polygons are always closed on the abscissa where the line extends from the lowest value with a frequency down to the abscissa. In this case, the rating of 1 had a frequency of zero, so the left side of the graph was already closed, but the rating of 11 had a frequency of 2, so an additional value above 11 and with a frequency of was needed in order to "close" the polygon. Similar to the polygon is the **line plot**, which is simply a frequency polygon that is not closed.



Line plots and histograms are also for presenting data from several groups within the same graph, that is, if several groups were measured on the same variable, these groups' data can be plotted within the same graph. For example, say I measure the time, in minutes, it takes students to complete an exam. I plot the data by sophomores and juniors, which are the two college classes in this particular course.



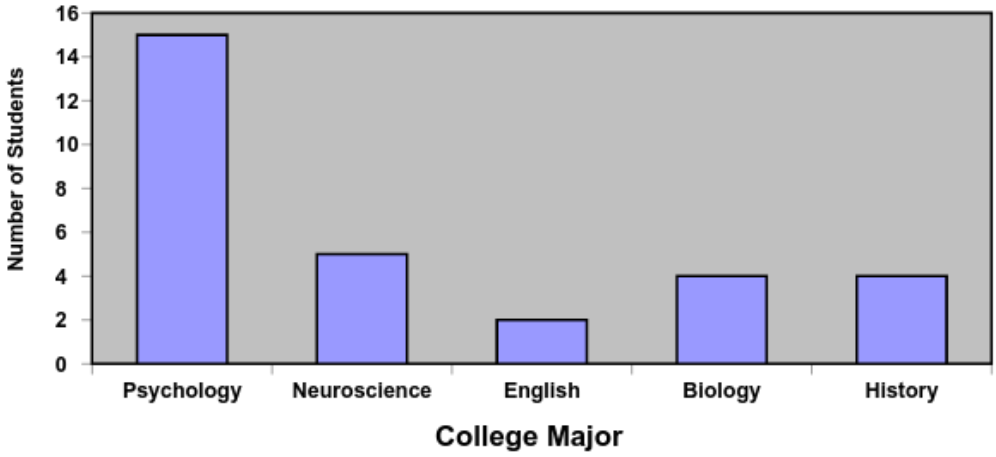
Here's a histogram showing the frequencies of sophomores and juniors who completed .



2.9 Graphs for Frequency Distributions of Qualitative Variables

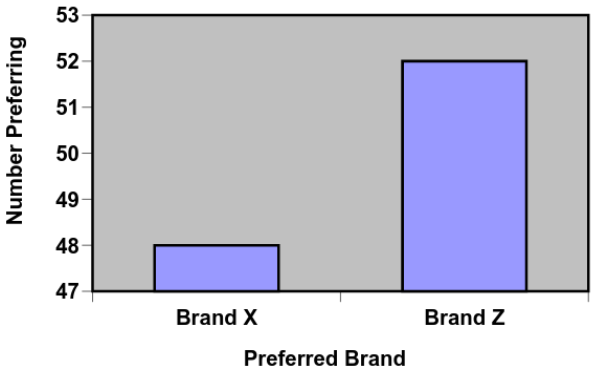
If the variable displayed on the abscissa is qualitative, a **bar graph** is used. Bar graphs and histograms look nearly identical, but histograms are used when the variable on the abscissa is quantitative and bar graphs are used when the variable on the abscissa is qualitative. Also, when using a histogram the bars

touch, but with bar graphs the bars do not touch. This is supposed to indicate a qualitative (categorical) difference between levels of the independent variable. An example of a bar graph of the data in Section 2.5 (college major) is below:

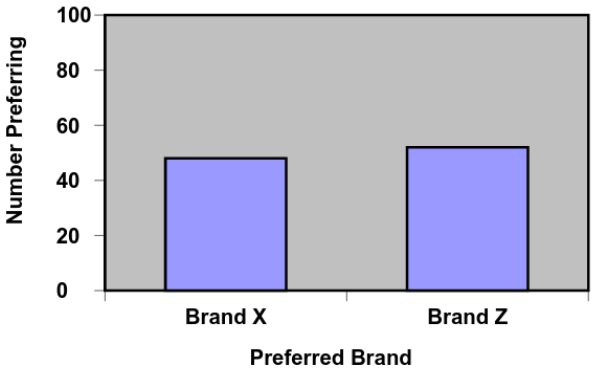


2.10 Misuse of Graphs

It is important to never misuse graphs, that is, don't use graphs to make a difference look bigger than it actually is or smaller than it is. This is one way people tend to lie with statistics. A classic example is when a researcher presents a **restricted range** of values on the y-axis to make a difference between levels on the x-axis look bigger than it really is. Specifically, a person may present only values around the range of their groups or conditions being measured in order to make any difference between the conditions look larger than it really is. For example, we have two types of coffee, Brand-X and Brand-Z. Brand-X is the best-selling coffee in the US and Brand-Z is a competitor. The company that produces Brand-Z does a taste test and counts the number of people that prefer Brand-X, and the number of people that prefer Brand-Z. The frequencies of people that like Brand-X and Brand-Z (out of 100 total people) are $f = 48$ and $f = 52$, respectively. The difference is very small and the correct way to present the data in a graph is below:



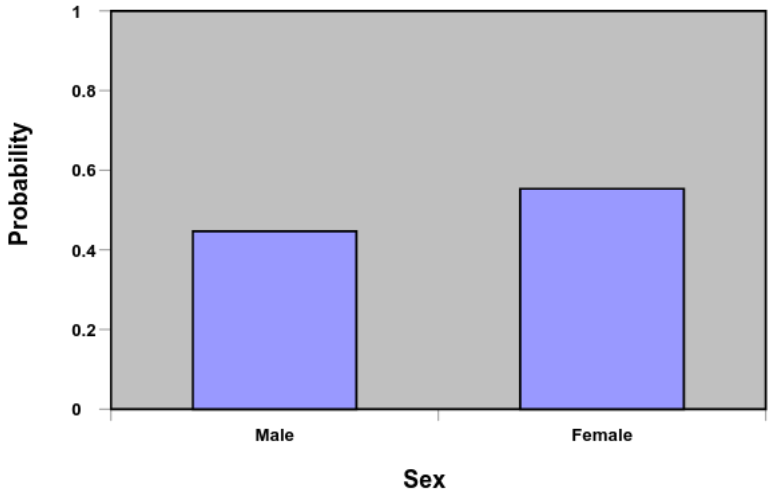
The difference in the heights of the two columns is not easy to see, which is good, because there is virtually no difference between the numbers of people who prefer Brand-X to Brand-Z. Below, is an example of how this data should not be presented graphically (but often is):



In this case the small difference in frequencies appears to be quite large, because the range on the ordinate has been restricted to values around the obtained frequencies. Don't do this! It is okay to restrict the range of values on the ordinate to *some* degree, but you should never blow a small difference out of proportion. Best advice: If it is a small difference, leave it that way!

2.11 Probability Distributions

A **probability distribution** is like the frequency distribution where the levels of one variable (e.g., Political Attitude) are listed with the frequency of each level. But a probability distribution is a graph that shows the probability on the y-axis as a function of all possible scores or categories of some variable on the x-axis. A probability distribution for a discrete variable is easy to create, because with a discrete variable, you should be able to determine the probability of being in one level of that variable compared to another. For example, say we wanted to construct a probability distribution for the independent variable sex (male/female) at the University of Scranton, based on fall 2011 data. That is, a probability distribution of full-time male and female undergraduate students at the University of Scranton. The x-axis shows each level of the independent variable *Sex* and the probability of male and female University of Scranton students is on the y-axis. This graph can be used to approximate the probability of selecting a male versus a female student attending the University of Scranton:



What if the variable on the x-axis was a continuous variable, like time? Technically speaking, a probability distribution for a continuous variable is not possible, because we could not determine the probability of every possible measurement, because the number of measurements is theoretically infinite. Instead, a probability distribution for a continuous variable is estimated by plotting a **probability density function**,

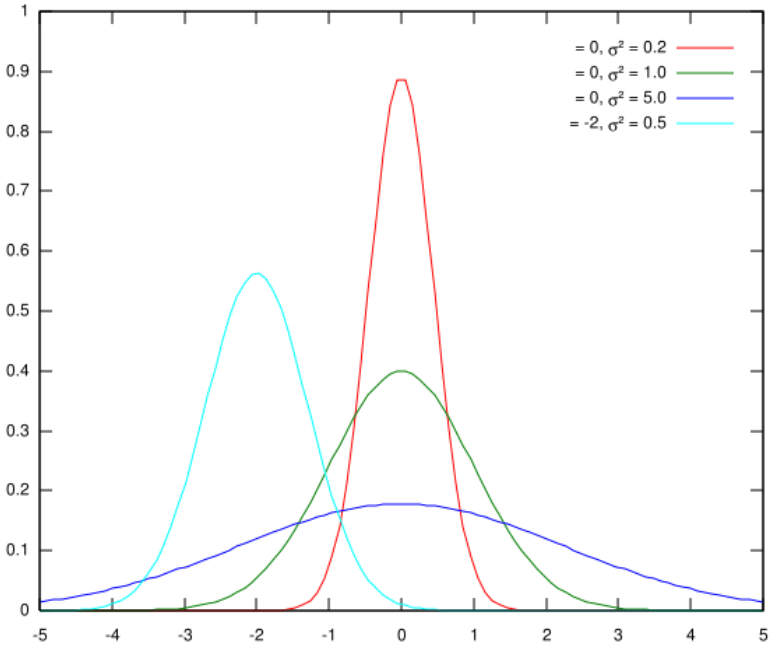
which is basically a function that approximates what a probability density function would look like for a set of parameters from a continuous variable.

One such distribution, the **normal distribution**, is a type of probability distribution that is estimated by means of a probability density function. The probability density function for a normal distribution is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Don't worry, I'm not going to make you memorize this function or use it; we simply do not have enough time to explain all of the terms and parameters in this function. The normal distribution is a **theoretical probability distribution** that is based off of this mathematical formula. All theoretical probability distributions are based off of mathematical assumptions and formulas, like that above.

The graph to the right includes several plots of normal distributions that were created with the function above (don't worry about what the various parameters mean right now). The y-axis is the probability and the x-axis includes values along some continuous variable.

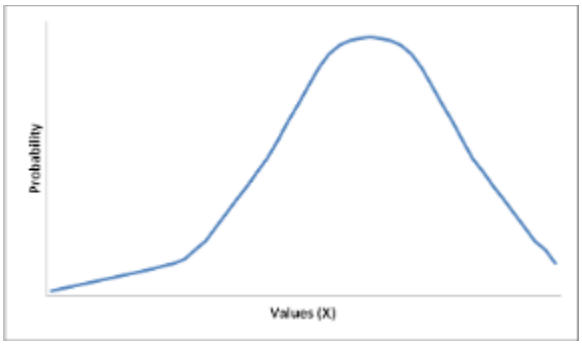


We will cover the normal distribution in detail later, but you should note a few things about the normal distribution. First, the total probability (area) under the curve is equal to 1. Second, the height of the curve at any point represents the probability of that value on the x-axis: The greater the height of the curve at any point, the more probable that value. As can also be seen, in the normal distribution the most probable values tend to be at the **center** of the distribution and the least probable values tend to lie in the **tails**, or ends, of the distribution. This is related to the concept of central tendency, which will discuss in Chapter 4. Finally, the normal distribution is perfectly symmetrical around its midpoint, so that the probability of being above that midpoint is .5 and the probability of being below that midpoint is .5.

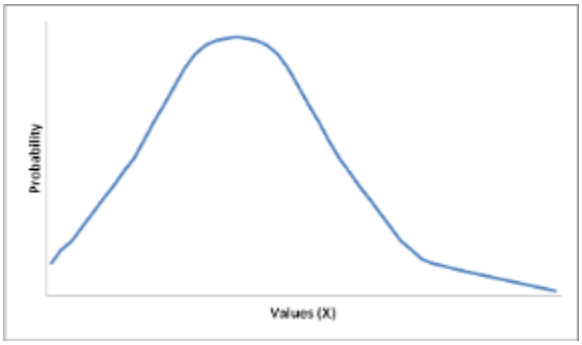
It is not always the case that a set of values is normally distributed; however, we often make a **normality assumption** with statistics. The normality assumption states that a distribution of collect scores is normally distributed. However, the normal distribution is theoretical and is based on mathematical assumptions. Probability distributions that are based on actually collected data are known as empirical **probability distributions** and often differ from normality. If the underlying distribution of scores is not normally distributed, it can throw off the results of a study. As long as a distribution of scores does not deviate too

much from normality there is no problem. When a distribution is not normal it can take on many different shapes and I have listed some of these below.

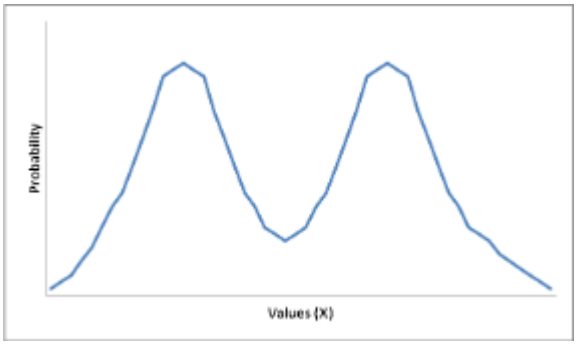
A **negatively skewed** distribution occurs when the hump lies to the right of center and the prominent tail lies to the left. This occurs when extremely small values of low probability occur in the data:



A **positively skewed** distribution occurs when the hump of the distribution lies to the left of center and the prominent tail lies to the right. This occurs when extremely large values of low probability occur in the data:



A **multi-modal** distribution occurs when there are two or more humps in the distribution. This usually occurs when you accidentally sample from two different populations:



Finally, a **rectangular distribution** occurs when each score in a distribution is equally frequent and the shape of the distribution resembles a rectangle:

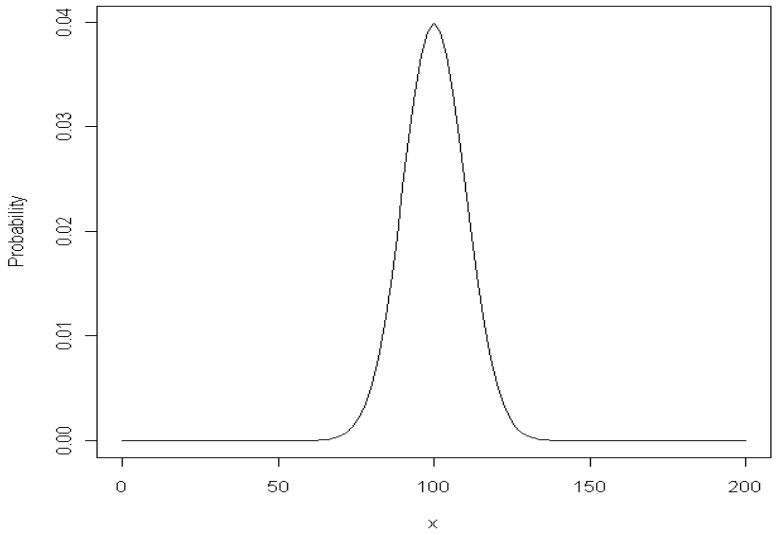
The latter two distributions rarely occur in data. The distributions we'll be concerned with most are the normal distribution, positively skewed distributions, and negatively skewed distributions.

2.12 Empirical vs. Theoretical Distributions

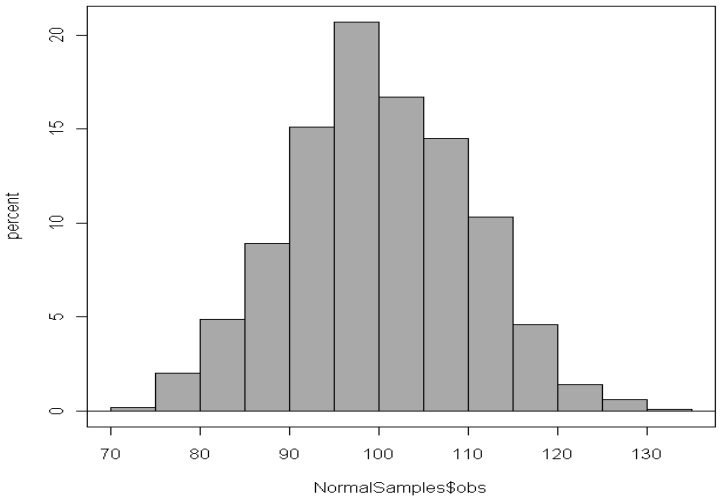
One final and important topic to mention is the difference between theoretical distributions and empirical distributions. **Empirical distributions** are based on actual measurements of a variable. Thus, empirical probability distributions express empirically observed relationships between variables. In contrast, **theoretical distributions** express relationships between variables mathematically based on a set of mathematical assumptions. The normal distribution discussed above is a theoretical probability distribution as it is not based on any collected data; rather, it is based on mathematical logic.

For example, below is a theoretical normal distribution with a mean of 100 and an empirical distribution derived from the theoretical distribution. Notice that the shapes are similar, but not identical.

Theoretical distribution:

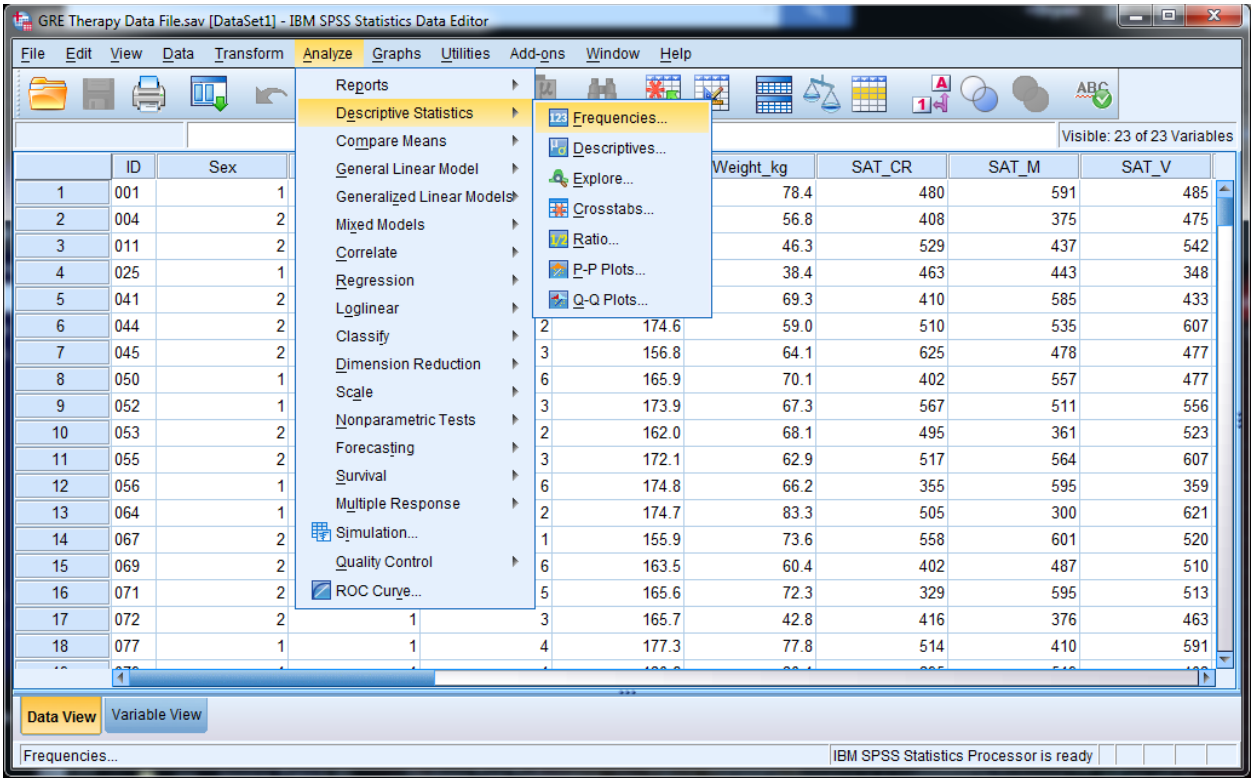


Empirical distribution:

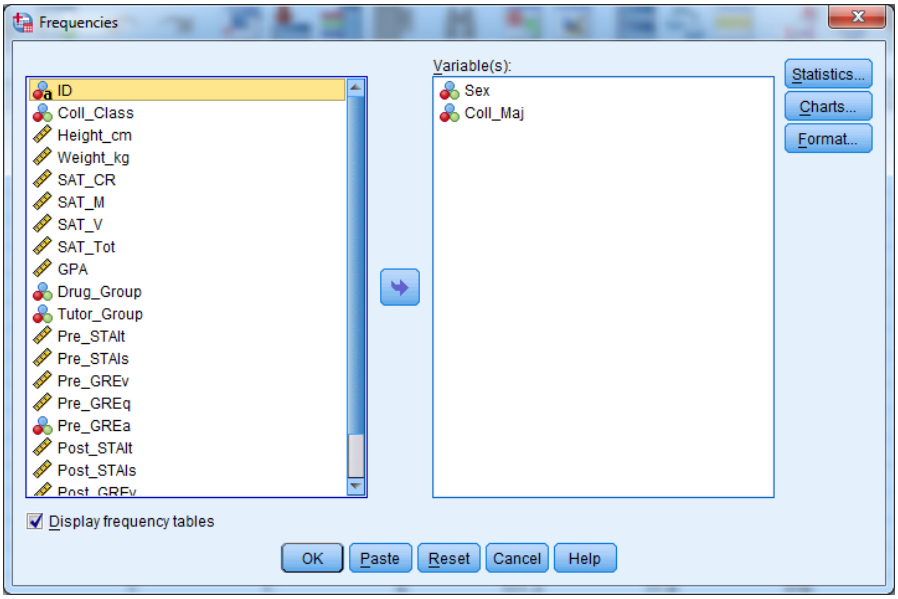


2.13 Frequency Distributions in SPSS

The following uses the GRE Therapy Data file. This data file is based on a hypothetical study examining the influences of a study-aid drug and types of tutoring on performance on the Graduate Record Examinations (GREs). To request frequency distributions in SPSS, from the Analyze menu, select Descriptive Statistics and the Select Frequencies (see below):



Assume we want to know the frequencies for each of the college majors (Coll_Maj) in the data set and the frequencies of the males and females (Sex); two qualitative variables. In the Frequencies window that opens, move Coll_Maj and Sex from the left to the right and click he OK button.



The resulting output is below, where it can be seen that there were 109 males (45.4% of the subjects) and 131 females (54.6% of the subjects). Additionally, you can see there were students from six different college majors in the data set, with each major represented roughly the same among the *n* = 240 students in the sample.

Frequency Table

Sex

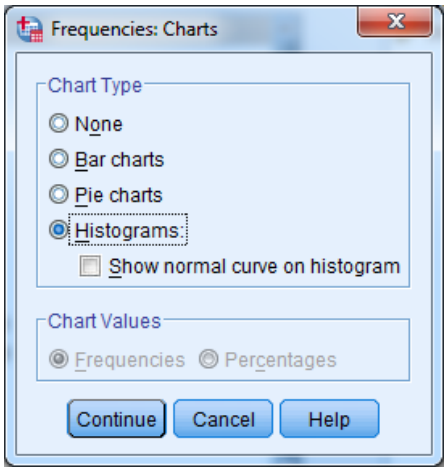
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Males	109	45.4	45.4	45.4
	Females	131	54.6	54.6	100.0
	Total	240	100.0	100.0	

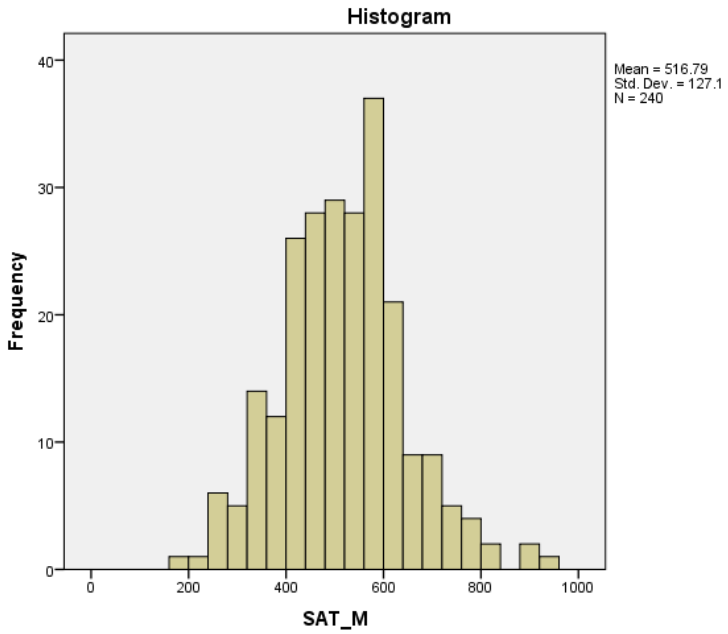
Coll_Maj

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Psychology	39	16.3	16.3	16.3
	History	40	16.7	16.7	32.9
	Biology	40	16.7	16.7	49.6
	Communications	30	12.5	12.5	62.1
	English	38	15.8	15.8	77.9
	Mathematics	53	22.1	22.1	100.0
	Total	240	100.0	100.0	

If you want to ask for frequencies from a quantitative variable, such as SAT Math scores (SAT_M), the procedure is the same; however, it is good to also request a histogram of the data (click the Charts button and select Histograms, see right), because quantitative variables can often take on many values and the

frequency of each value will be low. The histogram will "group" adjacent values together to show you the frequencies is similar values, making it much easier to interpret the frequencies (see below).





CH 2 Homework Questions

1. Identify whether each measure is a nominal measure, ordinal measure, interval measure, or ratio measure. Explain the reasons for your choices.

a. A cognitive scientist measures the time to solve a puzzle in seconds.
b. A historian groups the books on his shelf based on category.
c. A professor lists the publications on his curriculum vita in alphabetical order based on last names of the authors.
d. Students are asked to rate their political attitude on a scale from -5 (liberal) to 5 (conservative).
e. The morning weather usually reports the temperature in degrees Fahrenheit.
f. A professor counts the number of students that come to his office each week.

2. Indicate whether each measure is a nominal measure, ordinal measure, interval measure, or ratio measure. Explain the reasons for your choices.

a. inches on a yardstick
b. Social Security numbers
c. dollars as a measure of income
d. order of finish in a car race
e. intelligence test scores

3. Indicate whether each measure is a nominal measure, ordinal measure, interval measure, or ratio measure. Explain the reasons for your choices.

a. The speed of a slap-shop made by a hockey player.
b. The movie titles in your DVD collection.
c. First through third place in a pie-tasting contest.
d. Temperature measured in degrees-kelvin.
e. Temperature measured in degrees Celsius.
f. The number of baseball cards in a collection.
g. A rating given the following question on the scale below.

How much do you agree with the republican party?						
Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree or Disagree	Somewhat Agree	Agree	Strongly Agree

h. Letter grades: A, A-, B, B+.
i. Your name.

4. Indicate whether each of the following variables is discrete or continuous.

a. Height
b. gross domestic product
c. happiness
d. grains of sand in a sandbox

5. Identify each of the following as a qualitative or a quantitative variable:

a. weight
b. religion
c. income
d. age
e. gender
f. eye color

6. *Use the following data to answer the questions, below:* A professor recorded grades from a 5-point quiz:

5	1	0	2	1	3	4	5	0	1	2	5
3	4	3	2	3	5	3	5	3	2	4	3

a. Create a frequency distribution with the scores provided above. Be sure to include the absolute (raw) frequencies, the relative f frequencies, the cumulative frequencies, the cumulative relative frequencies, and the cumulative percentages.
b. What proportion of students had a score of 3?
c. What proportion of students had a score of 3 or less?
d. What proportion of students had a score less than 3?
e. What proportion of students had a greater than 3?
f. What proportion of students had a score of 3 or more?
g. How many students had a score of 4?
h. How many students had a score of 4 or less?
i. How many students had a score of 4 or more?
j. How many students had a score less than 4?
k. What is the probability that a randomly selected student has a score of 5?
l. What is the probability that a randomly selected student has a score of 5 or of 4?

7. *Use the following data to answer the questions, below:* A professor keeps records for how many times each of his 20 students come to his office hours each semester. The number of days each of these 20 students came to his offices hours are:

7	7	6	4	3
6	3	7	6	6
4	6	6	6	7
5	6	8	7	5

a. Create a frequency distribution with the data provided above. Be sure to include the absolute (raw) frequencies, the relative f frequencies, the cumulative frequencies, the cumulative relative frequencies, and the cumulative percentages.
b. What proportion of students came to the office 8 times?
c. What proportion of students came to the office less than 8 times?
d. What proportion of students came to the office 8 times or less?
e. What proportion of students came to the office more than 8 times?
f. What proportion of students came to the office 5 times or 6 times?
g. What proportion of students came to the office 6 times or more?
h. What proportion of students came to the office more than 6 times?
i. How many students did not come the office 4 times?
j. How many students came the office 4 times or less?
k. How many students came the office less than 5 times?

8. *Use the following data to answer the questions, below:* Dr. Evil owns a potato-chip company. He keeps records of how many days each of his employees were sick during the previous year. The number of days each employee was sick are:

8	7	8	9	6	5	0	1	2	3	3	4	5
6	8	6	4	6	0	9	2	3	4	7	6	8
5	4	0	1	2	3	5	6	6	4	5	3	2
0												

a. Construct a frequency distribution of these scores. Be sure to include the absolute (raw) frequencies, the relative f frequencies, the cumulative frequencies, the cumulative relative frequencies, and the cumulative percentages.
b. What proportion of employees were sick for exactly 6 days?
c. What proportion of employees were sick for 6 days or fewer?
d. What proportion of employees were sick for 6 days or more?
e. What is the probability an individual was sick for exactly 4 days?
f. What is the probability an was sick for 1 day or 0 days?

9. A college class is normally made up of many different college majors. You want to know which majors are taking an 'Introduction to Basket weaving' course. You ask each student in the class what their primary major is, and you find that the students in the class are Biology majors (B), English majors (E), Psychology majors (P), or Undeclared majors (U). The data for the students are as follows:

U U U U U E U B U P P E U U P B U
U E U B U U E U U B U U E U E P U
U E E P U U U P B U U U B U U E

Construct a frequency table that contains absolute frequencies, relative frequencies, and percentages of the data above.

10. *Use the following data to answer the questions, below:* In a recent study, the numbers of students from each of the following majors participated:

Biology = 13	Exercise Science = 12	Communication = 7
Undecided = 9	Counseling = 4	Occupational Therapy = 3
Education = 11	Computer Science = 2	Business = 8

History = 2 Philosophy = 2 Nursing = 4
Psychology = 2 Criminal Justice = 6

a. Construct a frequency distribution of these score and include the relative frequencies (rf) and percentages.
b. What proportion of students were Education Majors?
c. What proportion of students were Undecided Majors?
d. What is the probability that a randomly selected student is a History Major?
e. What is the probability a randomly selected student is a Counseling Major?
f. What is the probability a randomly selected student is a Psychology Major or a Computer Science Major?

11. What is a probability distribution? Why is the nature of probability distributions for qualitative and discrete variables different from that of continuous variables?

12. What is the difference between an empirical and a theoretical distribution?

13. A researcher surveyed 2000 individuals on their attitudes toward setting a national speed limit on interstate highways. One of the questions asked the individuals to indicate whether they thought that the national speed limit should be set at 65 mph. Responses could range from 1 ("definitely should not be set to 65 mph") to 4 ("definitely should be set to 65 mph"). Using the percentages below, compute the absolute (raw) frequencies, relative frequencies, cumulative frequencies, and cumulative relative frequencies for the set of responses below:

Response	*%*
4	35.0 %
3	15.0 %
2	20.0 %
1	30.0 %

14. The table below lists the numbers of majors enrolled in a Fundamentals of Psychology class. Using this frequency distribution table, create a bar graph (be sure to use labels).

Major	**f**
Undeclared	31
Communications	6
English	5
Education	11
Biology	7
Criminal Justice	3
Political Science	2
Business	3
Occupational Therapy	7
Psychology	8

15. *Use the following data to answer the questions, below:* A survey asked each participant to rate how much they agreed with the Democratic Party on a scale ranging from -3 to +3. The frequencies of each rating given for N = 177 participants are listed below.

Rating	-3	-2	-1	0	+1	+2	+3
Frequency	5	19	18	43	39	49	4

a. What type of data is "agreement with the Democratic party" being measured on?

b. What type of a graph should you use to plot these frequency data?
c. Construct the appropriate graph of the data above.

16. Here are frequencies of Freshmen, Sophomores, Juniors, and Seniors enrolled in a Statistics class:
Freshmen = 4 Sophomores = 19 Juniors = 5 Seniors = 3

a. What type of measurement scale is "class" measured on?
b. What type of graph (bar graph, line graph, or histogram) should you create to display these data?
c. Construct the appropriate graph of the data above.

Use the frequency distribution table below to complete exercises 17 – 21:

Score	f	rf	cf	crf
5	5	.2	25	1.00
4	3	.12	20	.8
3	7	.28	17	.68
2	5	.2	10	.4
1	3	.12	5	.2
0	2	.08	2	.08

17. Create a histogram for the absolute (raw) frequencies.

18. Create a line plot for the absolute (raw) frequencies.

19. Create a line plot for the relative frequencies. Is the shape similar or different to the graph in #12?

20. Create a line plot for the cumulative frequencies above.

21. Create a polygon for the cumulative relative frequencies above.

22. A principal in a small school measured the intelligence of fifth-grade students in her school. The intelligence test scores for those students were as follows:

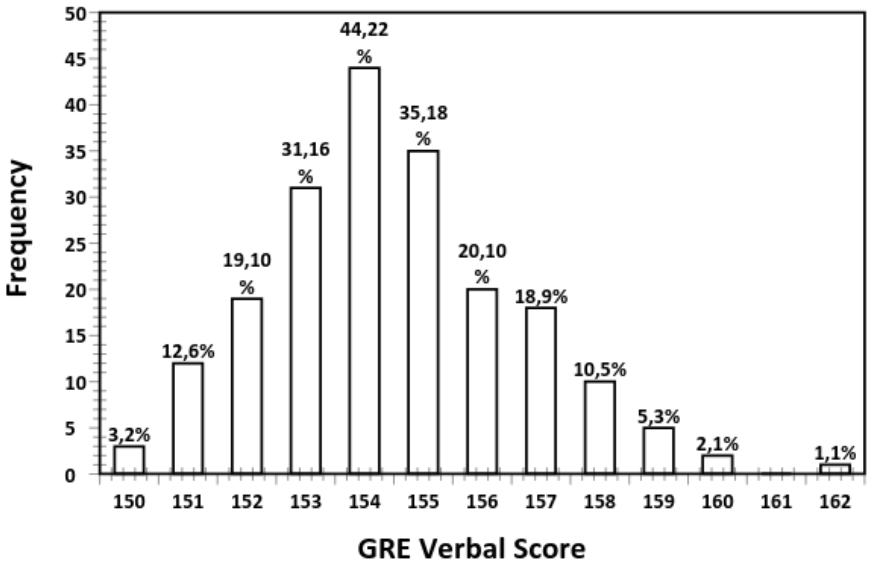
129	99	98	113	103	128	102	110	80	105
93	98	109	109	100	111	106	96	108	90
104	94	92	119	127	89	95	92	105	108
83	100	107	106	101	118	84	119	105	111
118	106	122	120	102	117	103	117	103	88

Draw a stem and leaf plot for the set of scores. Rotate the plot so the tens and hundreds digits appear at the base rather than on the right.

23. The IQ (intelligence) scores of thirty-three students from the same third grade classroom are presented. One set is for boys and the other set is for girls. Create a back-to-back stem-and-leaf plot so that you can visually compare the IQ scores of the boys with the IQ scores of the girls.

Boys:	113	113	122	116	122	134	134	105	122	126	129	133
	122	133	108	120	144	104						
Girls:	113	102	119	114	108	111	122	113	125	118	96	112
	124	122	115									

24. The following histogram shows the GRE Verbal scores for 200 students entering a local university. Use this graph to answer the questions that follow. (The numbers above each bar are the frequencies.)



a. How many of the students had scores between 152 and 156 (inclusive)?
b. How many students had scores less than 155?
c. How many students had scores greater than 160?
d. Do you agree with the following statement? “The distribution looks bell-shaped, with one outlier.”

Chapter 3: Rankings in a Distribution

Score	Math Percentile	Verbal Percentile	Score	Math Percentile	Verbal Percentile
800	94	99	500	26	60
780	89	99	480	23	54
760	85	99	460	20	48
740	80	99	440	17	43
720	75	98	420	14	37
700	70	97	400	12	31
680	66	95	380	10	25
660	61	93	360	8	20
640	57	91	340	6	15
620	52	88	320	5	10
600	47	85	300	3	5
580	42	81	280	2	3
560	38	76	260	2	1
540	34	70	240	1	1
520	30	65	220	1	

3.1 Percentiles and Percent Ranks

Frequency distributions are used to organize data, but can also be used to determine the relative location of a particular value within a distribution, that is, what proportion of scores are below or above a value. The **percent rank** (**percentile rank**, **PR**) is the percentage of scores less than or equal to a value (X) in a distribution. The percent rank is similar to the cumulative relative frequency of a value and, similarly, the **percentile** (X_P) is the value (X) a certain percentage of scores are less than or equal to; hence, a percentile is the value associated with a specific percentile rank.

You've probably heard these terms before when discussing how well you performed on a standardized test such as the SATs. For example, when you get your SAT score you are told your percentile rank, which is the percentage of test-takers who you scored greater than or equal to on the SATs. Here's a portion of the raw SAT scores and percentile ranks in each of the three SAT sub-tests:

Score	***Critical Reading***	***Mathematics***	***Writing***
560	69	63	72
550	66	61	69
540	63	58	66
530	60	55	63

In the table above, the raw scores in the left column would be considered percentiles and the values under each subtest column are the percentile ranks associated with each raw score. For example, if you took the SATs and received a 540 on the Mathematics test, your percentile rank is about 58%. This means that your score is greater than or equal to about 58% of all test-takers. Similarly, if you scored 560 on the Writing test, you scored greater than or equal to about 72% of all test-takers. Thus, percentile ranks tell you how well a score is relative to the entire distribution.

Alternatively, if you wanted to know approximately what score on the Critical Reading test is greater than or equal to 66% of all other scores, you look up a percentile rank of 66% under the Critical Reading column, then find the value in the Score column (550); this would be the approximate score that is greater than or equal to 66% of all Critical Reading scores. Similarly, if you wanted to know what score is greater than or equal to 65% of all Mathematics test scores you would locate the percentile rank in the Mathematics column

that is immediately greater than 65%, which is 66%. The score associated with that percentile rank (540) will be the percentile that is approximately greater than or equal to 66% of all other Mathematics scores. The percentiles and percentile ranks above are *approximate*, because, in practice you must calculate the actual percentile or percentile ranks, which is covered in the following sections.

3.2 Calculating Percent Ranks

For the following examples of calculating percentiles and percent ranks, we'll use the frequency distribution for quantitative data from Chapter 2, which I have reproduced below:

Political Attitude (X)	f	rf	%	cf	crf	c%
11	2	.04	4%	50	1.00	100%
10	0	.00	0%	48	.96	96%
9	4	.08	8%	48	.96	96%
8	6	.12	12%	44	.88	88%
7	6	.12	12%	38	.76	76%
6	13	.26	26%	32	.64	64%
5	10	.20	20%	19	.38	38%
4	2	.04	4%	9	.18	18%
3	4	.08	8%	7	.14	14%
2	3	.06	6%	3	.06	6%
1	0	.00	0%	0	.00	0%

Recall, the percentile rank for a value is the percentage of scores a value is greater than or equal to. The formula for determining percentile rank for a value is:

$$PR_X = \left[\frac{n_L + (n_w / i)(X - L)}{N}\right] 100$$

We first need a value to determine a percentile rank. Say we want to find the percentile rank for a Political Attitude value of X = 8. Locate that value in the frequency distribution table (the row for X = 8 is highlighted) and make this row your 'reference point'. The values for each term in the percentile rank formula above are based off this reference point. The following are descriptions of each term in the formula, along with where to find the value in the frequency distribution table:

n_L is the number of scores less than the value we are calculating the percentile rank for, that is, the cumulative frequency (cf) of X = 7. The cumulative frequency of X = 7 is 38, so $n_L = 38$. **n_W** is the number of scores associated with the value we are calculating the percentile rank for, that is, the frequency (f) of X = 8. The frequency of X = 8 is 6; hence $n_W = 6$. **i** is the number of values associated with the percentile rank, which in this example is i = 1 (i.e., X = 8). In some distributions i can be greater than 1; but we do not cover distributions. **L** is the lower real limit (LRL) of the value we are calculating the percentile rank for. The lower real limit is one-half unit below our measured value, so the lower real limit of 8 is L = 7.5. **X** is the value for which we are calculating the percentile rank (X = 8). **N** is the total number of scores in the distribution, which in this case is 50. Once you have located each necessary value, plug them into the formula and solve for the percentile rank of X = 8.

$$PR_X = \left[\frac{38 + (6/1)(8 - 7.5)}{50}\right] 100 = 92\%$$

Thus, X = 8 is greater than or equal to exactly 92% of all other scores in the set of data. Or, another way to state this is that 92% of the individuals surveyed self-reported a political attitude that was at most 'somewhat conservative'.

3.3 Calculating Percentiles

Recall, a percentile is the value greater than or equal to a certain percentage of scores. Below, I reproduced the frequency table from above. Let's say we want to determine the percentile (X) associated with a percentile rank of 35%, that is, what value is greater than or equal to 35% of the scores?

Political Attitude (X)	f	rf	%	cf	crf	c%
11	2	.04	4%	50	1.00	100%
10	0	.00	0%	48	.96	96%
9	4	.08	8%	48	.96	96%
8	6	.12	12%	44	.88	88%
7	6	.12	12%	38	.76	76%
6	13	.26	26%	32	.64	64%
5	10	.20	20%	19	.38	38%
4	2	.04	4%	9	.18	18%
3	4	.08	8%	7	.14	14%
2	3	.06	6%	3	.06	6%
1	0	.00	0%	0	.00	0%

Like calculating percentile ranks in section 3.2, we need to determine a reference point in the frequency distribution table. This reference point is based on where the percentile rank (35%) falls within the cumulative percentages. As you will notice, 35% does not appear in the cumulative percentage column. In this case, locate the cumulative percentage greater than the percentile rank we are looking for, which in this case is 38%. This is the reference point, which is highlighted in green. The formula for calculating percentile rank for a value is:

$$X_P = L + \left[\frac{(N)(P) - n_L}{n_W}\right] i$$

Each term in this formula are the same as those found in the formula for calculating percentile ranks in Section 3.2. The only new term is **P**, which is the percentile rank we are interested in (35%), expressed as a proportion (.35). The remaining values from the table above have been inserted into the formula to the right. In this example, X = 4.85 is greater than or equal to 35% of all of the scores in this distribution. Or, the percentile that is associated with 35% of the scores is 4.85.

$$X_P = 4.5 + \left[\frac{(50)(.25) - 9}{10}\right] 1 = 4.85$$

CH 3 Homework Questions

Use the following information to complete Exercises 1 – 3: A statistics professor gave a 5-point quiz to the 50 students in his class. Scores on the quiz could range from 0 to 5: The following frequency table resulted:

Quiz Score	f	rf	cf	crf	c%
5	4	.08	50	1.00	100%

4	10	.20	46	.96	96%
3	14	.28	36	.72	72%
2	10	.20	22	.44	44%
1	8	.16	12	.24	24%
0	4	.08	4	.08	8%

1. Compute the values that define the following *percentiles*:

a. 25th b. 50th c. 55th d. 75th e. 80th f. 99th

2. What is the interquartile range of the data in #1?

3. Compute the exact *percentile ranks* that correspond to the following scores:

a. 2 b. 3 c. 4 d. 1

Use the following information to complete Exercises 3 – 5: A political scientist took a pool of the political attitudes of the students in one of his classes. Students were asked to rate, on a scale from 1 to 11, "What is your overall political attitude?", where 1 = extremely liberal and 11 = extremely conservative. The following frequency analysis resulted:

Political Attitude Score	f	rf	cf	crf
11	1	.015	67	1.000
10	3	.045	66	.985
9	14	.209	63	.940
8	6	.090	49	.731
7	2	.030	43	.642
6	10	.149	41	.612
5	9	.134	31	.463
4	3	.045	22	.328
3	11	.164	19	.284
2	7	.104	8	.119
1	1	.015	1	.015

4. Compute the values that define the following *percentiles*:

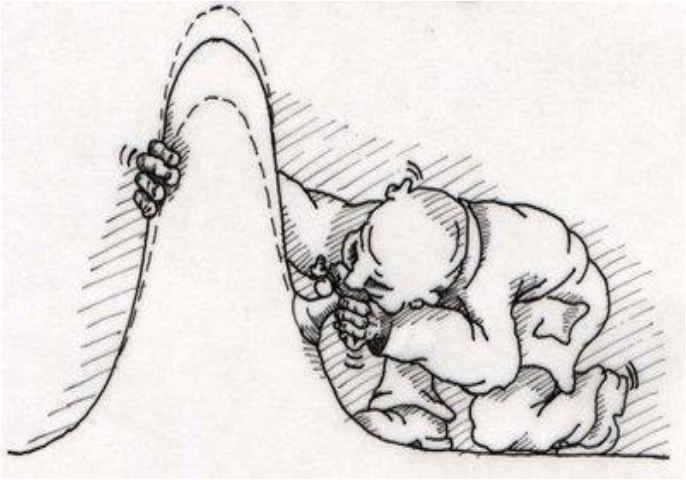
a. 25th b. 50th c. 57th d. 75th

5. What is the interquartile range of the data in #4?

6. Compute the exact *percentile ranks* that correspond to the following scores:

a. 3 b. 5 c. 7 d. 9

Chapter 4: Central Tendency

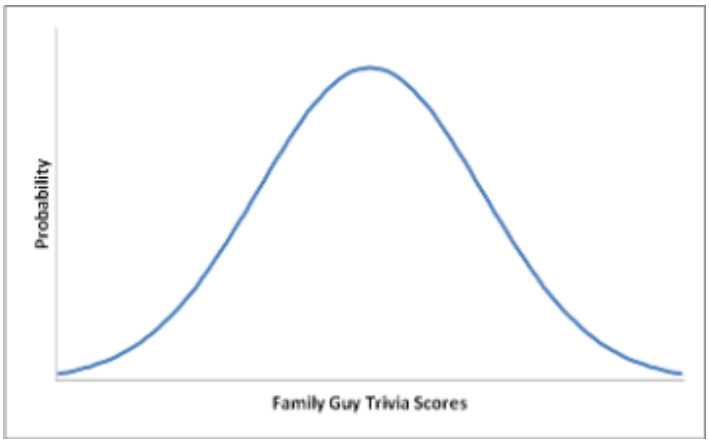


4.1 Measures of Central Tendency

So you've collected some data, created a frequency distribution, and made a graph of your data...now what? The statistical procedures from chapters 1 and 2 dealt with organizing data, but even after organizing data, it's still complex and it would be nice to have a single value to represent a data set. A **descriptive statistic** is a such value that describes and represents a set of data. But what value from a distribution should be used as the representative and why is any one value in the data set better than another value? There are three characteristics you need for a god descriptive statistic.

First, a good descriptive statistic should be similar to many scores in a distribution; hence, its value should have a high frequency. Second, it should balance the distribution so some scores are greater than it and some scores are less, so neither the greater-than or less-than scores are overrepresented in the descriptive statistic. Finally, a good descriptive statistic should take individual values from the distribution into account so no value is left out.

Examine the hypothetical distribution of Trivia Scores to the right. Remember, in a normal distribution the most frequent scores cluster near the center and less frequent scores fall into the tails. **Central tendency** simply means most scores in a normally distributed set of data tend to cluster near the center of a distribution. Thus, the central tendency area is associated with most of the scores in a set of data; about 68% of the scores to be exact.



Recall, a characteristic of a good descriptive statistic is it should be similar to many scores; hence, values near the center of a distribution make good descriptive statistics, because those values are similar to many values in the data. Also, because a normal distribution is symmetrical around its center with half of the scores above the center and the other half of the scores below the center, a score from this central tendency area should roughly equally represent the larger values and the smaller values of the distribution. Thus, **measures of central tendency** are useful for representing a

set of data, because measures of central tendency describe where most of a distribution lies along some continuum.

4.2 The Mode

This section uses the following set of *n* = 50 quiz scores (X), which have a range of 0 - 10:

10 10 10 9 9 9 9 8 8 8 8 8 7 7 7 7 7 7 6 6 6 6 6 6 6
5 5 5 5 5 5 5 5 5 4 4 4 4 4 4 4 4 3 3 3 3 3 2 2 1

The **mode** (**Mo**) is the most frequently occurring score in a data set. IN the data set above, X = 5 has the greatest frequency (*f* = 9); hence, the mode is 5.

The mode is easy to determine, but it has limitations as there can be more than one mode in a data set. Recall from Chapter 2, a theoretical normal distribution has one hump; hence, a normal distribution has one mode. However, an empirical distribution can be **multi-modal**, meaning there is more than one hump (more than one mode). For example, in the data below there are two most frequent scores, X = 9 and X = 4, both with frequency of 6. In this case, both 9 and 4 are the modes, but neither one is better than the other:

10 10 9 9 9 9 9 9 8 8 7 7 6 6 5 5 4 4 4 4 4 4 3 3 2 2 1 1

Another problem with the mode is every score can be a mode. This occurs when you have a **rectangular** or **uniform distribution** and each score in the distribution is equally frequent. For example, the data in the set below has one of each value and each value is a mode:

10 9 8 7 6 5 4 3 2 1 0

Another problem is that the mode might be at one end of a distribution, not at the center. In the data below, the mode of 10 represents the larger values quite well but not the smaller values.

10 10 10 10 10 10 10 9 8 7 6 5 4 3 2 1 0

Finally, even if the scores in a set of data are normally distributed it is possible that the most frequent score has a frequency that is only one or two counts greater than the second most frequent score. In the first example above the mode was X = 5. However, X = 4 had a frequency of eight, which is only one less than the frequency of nine for X = 5. Why is 5 a better mode than 4? There is no answer.

4.3 The Median

The **median** (**Md**) is the middle score of a distribution, because 50% of the scores in a distribution are greater than the median and 50% are less than the median; hence, the median is the 50th percentile. This generally makes the median a better measure of central tendency than the mode, because the median balances a distribution such that equal numbers of scores are greater than and less than the median.

To find the median value of a distribution you must determine the **median's location** and then find the value at that location. Determining median location and value differs whether you have an even or an odd number of scores. For example, in the set of scores below there are *n* = 11 scores; hence, an odd number of scores:

10 10 9 7 7 6 5 4 3 2 2

When you have an odd number of scores, use the following equation to identify the median location:

$$Md = (N + 1)/2$$

The N is the number of scores in the data set (11). Solving for the median location, we get:

$$Md = (11 + 1)/2 = 6\text{th}$$

This indicates the median is located at the sixth position in the distribution. To determine the median value, make sure the scores are rank-ordered from smallest to largest. Next, count off six positions starting with the smallest value. Once you find the sixth position the value at that position is the median value. In this case the median value is X = 6.

X:	*2*	*2*	*3*	*4*	*5*	*6*	*7*	*7*	*9*	*10*	*10*
Ordinal Position:	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th

If there is an <u>even</u> number of scores, you must find the two positions surrounding the median, identify the values at those two locations, and take the average of those values. Hence, with an even number of scores, the median is the average of the two scores that surround the median's position. The following data an even number of scores:

20 19 18 16 15 14 12 11 11 11 10 9

Use the following to determine the two positions around the median:

$$Md = N/2 \text{ and } (N + 2)/2$$

Again, N is the number of scores in the data set (12). Solving for the median locations we have:

$$Md = 12/2, (12 + 2)/2 = 6\text{th}, 7\text{th}$$

The median is between sixth and seventh positions. To determine the median value make sure the scores are rank-ordered from smallest value to largest value. Next, starting with the smallest value, count off six positions and seven positions. Once you locate the sixth position and seventh position take the average of the values at those two positions. This will be the median value:

X:	9	10	11	11	11	12	14	15	16	18	19	20
Ordinal Position:	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	20th

In this example, X = 12 is at the sixth position and X = 14 is at the seventh position. The average of 12 and 14 is (14+12)/ = 13, which is the median value.

The median does have limitations. The median does not take into account the actual values of the scores in a set of data. For example, take the following two sets of scores, each with *n* = 5 scores:

Set 1:	1	4	5	7	10
Set 2:	4	4	5	6	6

In each set the median is equal to 5; hence, the median value is unaffected by the values greater and less than its value. In order to account for the individual values in a distribution, which makes is a desirable characteristic of a good measure of central tendency; we must calculate the **mean** of a distribution, which we will get to shortly.

Although the median does not take individual values into account, it is frequently reported as a measure of central tendency when data sets are incomplete or data are severely skewed; a point we return to later. One example of the median being used instead of the mean is median household income. The median is reported as a measure of central tendency for household income, because there is extreme positive skew in household income data; that is, a low frequency of extremely high incomes, which pull the mean up.

4.5 Quartiles

The median divides a distribution in half, but a researcher may want to divide a distribution into smaller parts, perhaps thirds, quarters, fifths, sixths. In particular **quartiles** split a distribution into fourths (quarters). There are actually three quartile points in a distribution: The 1st quartile (Q1) separates the lower 25% of the scores from the upper 75%; the 2nd quartile (Q2) is the median and separates the lower 50% of the scores from the upper 50%; and the 3rd quartile (Q3) separates the upper 25% of the scores from the lower 75%.

Determining the quartile value is like determining the median. First, determine the locations of the first and third quartiles. Second, locate the values at those locations. Also, like when determining the median, there are different procedures for determining the quartiles based on your having an even versus an odd number of scores.

For the 1st quartile (Q_1), if you have an odd number of scores, use the following: $Q_1 = (N+1)/4$
For the 1st quartile (Q_1), if you have an even number of scores, use the following: $Q_1 = (N+2)/4$
For the 3rd quartile (Q_3), if you have an odd number of scores, use the following: $Q_3 = (3N+3)/4$
For the 3rd quartile (Q_3), if you have an even number of scores, use the following: $Q_3 = (3N+2)/4$

One thing to note is the value you calculate from each expression above provides you with the position of the quartile. This location may be a whole number (2, 3, 4, 5, etc.) and indicates the value at that location is the first or third quartile. However, the location can also end in a .5 (1.5, 2.5, 3.5, 4.5, 5.5, etc.), which indicates the quartile is the average of two values. For example, if you calculate the first quartile position and it comes out as $Q_1 = 2.5$, this tells you the first quartile is the average of the values at the second and third positions, because the '2.5th position' is between the second and third positions. For example, the following set of n = 9 scores include an odd number of scores:

7 7 8 9 10 11 11 14 15

The locations of the first and third quartiles are:

$$Q_1 = (9+1)/4 = 2.5\text{th position}$$

$$Q_3 = (3 \times 9 + 3)/4 = (27 + 3)/4 = 7.5\text{th position}$$

Below, each score is listed with the ordinal position of each score from the data set above.

Position:	1st	2nd	3rd	4th	5th	6th	7th	8th	9th
Score:	7	7	8	9	10	11	11	14	15

$Q_1 = 2.5$ means the first quartile is the average of the values at the second (X = 7) and third (X = 8) positions; hence, the value of the first quartile is (7 + 8)/2 = 7.5. The $Q_3 = 7.5$ means the third quartile is the average of the values at the seventh (X = 11) and eighth (X = 14) positions; hence, the third quartile is (11 +14)/2 = 12.5. The difference between the third quartile and the first quartile (Q3 - Q1) is the **interquartile range**, which in this case is equal to, 12.5 - 7.5 = 5. The interquartile range is the middle 50% of the scores.
As a second example, the following set of n = 12 scores include an even number of scores:

6 7 7 8 9 10 10 11 11 14

The locations of the first and third quartiles are:

$$Q_1 = (10+2)/4 = 3\text{rd position}$$

$Q_3 = (3 \times 10 + 2)/4 = (32 + 2)/4 = 8$th position

Below, each score is listed with the ordinal position of each score from the data set above.

Position:	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Score:	6	7	7	8	9	10	10	11	11	14

The $Q_1 = 3$ means the first quartile is at the third position, which is X = 7. Similarly, the $Q_3 = 8$ means the third quartile is at the eighth positions, which is X = 11. The interquartile range is Q3 - Q1 = 11 - 7 = 4.

4.6 The Mean

You have probably seen the symbol Σ (Greek **sigma**), which in mathematics is the **summation** operator and means to add all values to the right of Σ. Often, a variable (X) is presented to the right of the Σ, which means you need to add up all values associated with the variable. For example, ΣX means to sum up all values that belong to variable X. Importantly, Σ is a grouping symbol like a set of parentheses, so you should do any mathematical work and operations to the right of Σ before adding. In the table to the right, five professors who teach at Whatsamatta-University are listed with their 2005 and 2006 salaries. We'll use this table of data to perform some operations using Σ.

Professor	**2005 Salary (Y)**	**2006 Salary (X)**
Dr. Java	39000	41000
Dr. Spock	43500	46000
Dr. Evil	48000	51000
Dr. Griffin	52500	56000
Dr. Who	57000	62000

From the data above, ΣX is equal to the sum of the 2005 salaries: ΣX = 39,000 + 43,500 + 48,000 + 52,500 + 57,000 = 240000. As another example, ΣY is equal to the sum of the 2006 salaries: ΣX = 41,000 + 46,000 + 51,000 + 56,000 + 62,000 = 256000, Lastly, Σ Dr. Evil is equal to the sum of Dr. Evil's salaries: Σ Dr. Evil = 48,000 + 51,000 = 99000.

What about if there are a set of mathematical operations on variables to the right of Σ? Remember, Σ is a grouping symbol like a set of parentheses and everything to the right of Σ must be completed before summing the resulting values. For example, say that we have the following set of data for variables X and Y:

X	Y
10	5
9	5
8	4
6	4
4	3
3	2

We want to find ΣX^2. What would you do? For ΣX^2 you first find the X^2 values before adding, because ΣX^2 is telling you to add up all of the X^2 values. To do this you take the square of each value associated with X and then add those squared values. Thus:

$$\sum X^2 = 10^2 + 9^2 + 8^2 + 6^2 + 4^2 + 3^2 = 100 + 81 + 64 + 36 + 16 + 9 = 306$$

Here are some other examples:

- $(\Sigma X)^2 = (10 + 9 + 8 + 6 + 4 + 3)^2 = (40)^2 = 1600$
- $\Sigma(X - Y) = (10 - 5) + (9 - 5) + (8 - 4) + (6 - 4) + (4 - 3) + (3 - 2) = 5 + 4 + 4 + 2 + 1 + 1 = 17$
- $\Sigma(X - Y)^2 = (10 - 5)^2 + (9 - 5)^2 + (8 - 4)^2 + (6 - 4)^2 + (4 - 3)^2 + (3 - 2)^2 = 5^2 + 4^2 + 4^2 + 2^2 + 1^2 + 1^2 = 25 + 16 + 16 + 4 + 1 + 1 = 63$
- $\Sigma XY = (10*5) + (9*5) + (8*4) + (4*3) + (3*2) = 50 + 45 + 32 + 24 + 12 + 6 = 169$

The **mean** is the arithmetic average of all the scores in a distribution. The mean is the most-often used measure of central tendency, because it evenly balances a distribution so both the large and small values are equally represented by the mean and also takes into account all individual values. It is this second point that is the main advantage of the mean: it accounts for individual scores in a data set.

To calculate the mean, add together all the scores in a distribution (ΣX) and then divide that sum by the total number of scores in the distribution (n). The expressions for calculating the mean from a sample and from a population are (in APA format, an italicized *M* is used to represent the mean):

$$\overline{X} = M = \frac{\Sigma X}{n}$$

$$\mu = \frac{\Sigma X}{N}$$

For example, we calculate the mean from the sample of n = 11 scores from the median example in Section 4.4. The scores in that set of data were:

2 2 3 4 5 6 7 7 9 10 10

To calculate the mean, determine the sum of the values ($\Sigma X = 65$). Next, divide that value by the total number of scores in the distribution ($n = 11$). Thus:

$$M = (65)/11 = 5.909$$

This value ($M = 5.909$) is close to the median value (Md = 6). The difference between the mean and the median reflects the mean's taking into account the individual values. To emphasize the fact that the mean takes into account all values in a set of data, recall the two sets of $n = 5$ scores from Section 4.4 that had the same median (Md = 5):

Set 1:	1	4	5	7	10	Md = 5
Set 2:	4	4	5	6	6	Md = 5

Calculating the mean for Set 1, we have $M = 27/5 = 5.4$, and for Set 2 we have $M = 25/5 = 5$. You should begin to see why the mean is a more accurate measure of central tendency as it takes the individual scores into account; the median does not.

The mean is defined as the **mathematical center** of a distribution. This means the positive deviations from the mean and negative deviations from the mean (i.e., differences between scores and the mean) will cancel each other out. For example, by subtracting the mean from each score in Set 2 from above and summing those values, we get:

X	*4*	*4*	*5*	*6*	*6*

M	5	5	5	5	5
(X - *M*)	-1	-1	0	1	1

The negative deviations (-1 + -1 = -2) and the positive deviations (1 + 1 = 2) cancel out (-2 + 2 = 0), such that Σ(X - *M*) = 0. Thus, the mean evenly balances large scores and small scores in the distribution, which is will be true in any distribution.

The sample mean is the preferred measure of central tendency, because it is the best **unbiased estimator** of the population mean (μ). This is because positive and negative differences between the sample's mean and the individual values in that sample will cancel out (i.e., the mean is a 'balancing point'). As long as the sample was randomly selected from and is representative of the population, the values that would be expected to be obtained from the population should be represented in the sample.

There is one problem with the mean and ironically it has to do with taking individual values into account. Although this is a good characteristic, by taking individual values into account the mean can be influenced by extremely large or extremely small values (**outliers**). If a few extremely large or extremely small values occur in a data set they can pull the mean in that direction and away from the center of the distribution. Specifically, extremely small values will pull the mean down and extremely large values will pull the mean up. This only occurs in skewed distributions. In such cases, it is good to use the median as a measure of central tendency, because although it does not take individual values into account the median is not influenced by individual values; hence, it cannot be influenced by extreme scores.

4.7 Central Tendency and Scales of Measurement

An important consideration is which measure of central tendency should be used for each of the various measurement scales (nominal, ordinal, interval, or ratio). When a variable is measured on a nominal scale, the mode is the best measure of central tendency. For example, if I measured the sex of each subject in a research study I might have 5 males and 15 females. The mode is the 15 females. I cannot calculate the "median sex" of the "mean sex" when I am counting people. When a variable is measured on an ordinal scale any one of the three measures of central tendency *could* be used although the median may be more appropriate, because the individual values on an ordinal scale are meaningless and are relative only to the placement of other values. Finally, when a variable is measured on an interval or ratio scale, any of the three measures of central tendency is appropriate; however, the mean is generally preferred, because the mean accounts for every score in the distribution.

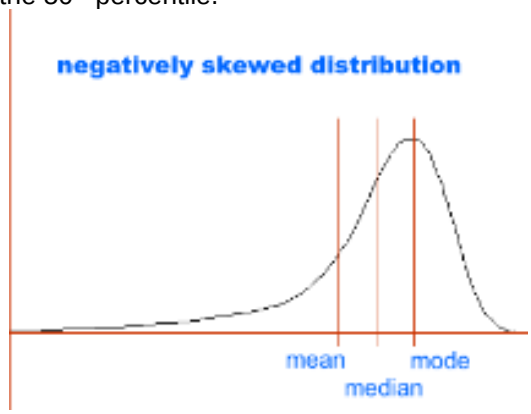
One issue with using the mean is whether a variable is discrete or continuous. If a variable is continuous then rounding the mean to any number of decimal places is perfectly appropriate, because data from a continuous variable can take on any value. For example, if I was measuring time in seconds, a mean of *M* = 1.25 seconds or of M = 6.8923 seconds make sense. In contrast, if a variable is discrete, the rounding is still appropriate, but remember, the fractions have to be carefully interpreted. For example, if I calculate the average number of students in all statistics classes, it could might be *M* = 28.75, but, I cannot have 0.75 of a person. This mean of 28.75 simply means the average number of students in statistics classes is between 28 and 29 students per class, but closer to 29 people.

4.8 The Mean Median and Mode in Normal and Skewed Distributions

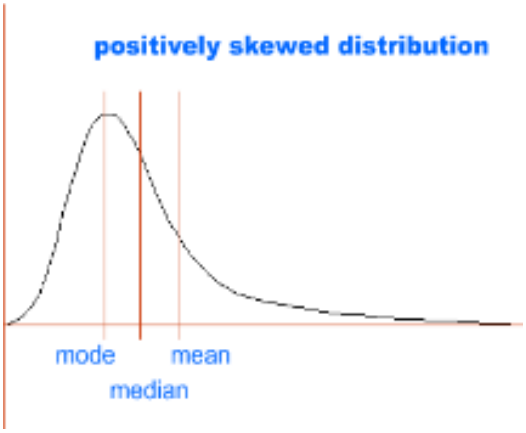
The positions of the mean, median, and mode are influenced by whether a distribution is normally distributed or skewed. If data are normally distributed, the mean is equal to the median, and the mean and median are equal to the mode. In a perfectly normal distribution, the data is symmetrical around the most frequent value at the center of the distribution. Because the distribution is perfectly symmetrical around the center, 50% of the scores must lie above the center point, which is also the mode, and the other 50% of the scores must lie below. Thus, the mode and the median must be equal. Finally, because the distribution is perfectly symmetrical the differences between the mean and the values larger than the mean must cancel

out all of the differences between the mean and the values less than the mean, such that the mean is also at the exact center (i.e., the median and the mode).

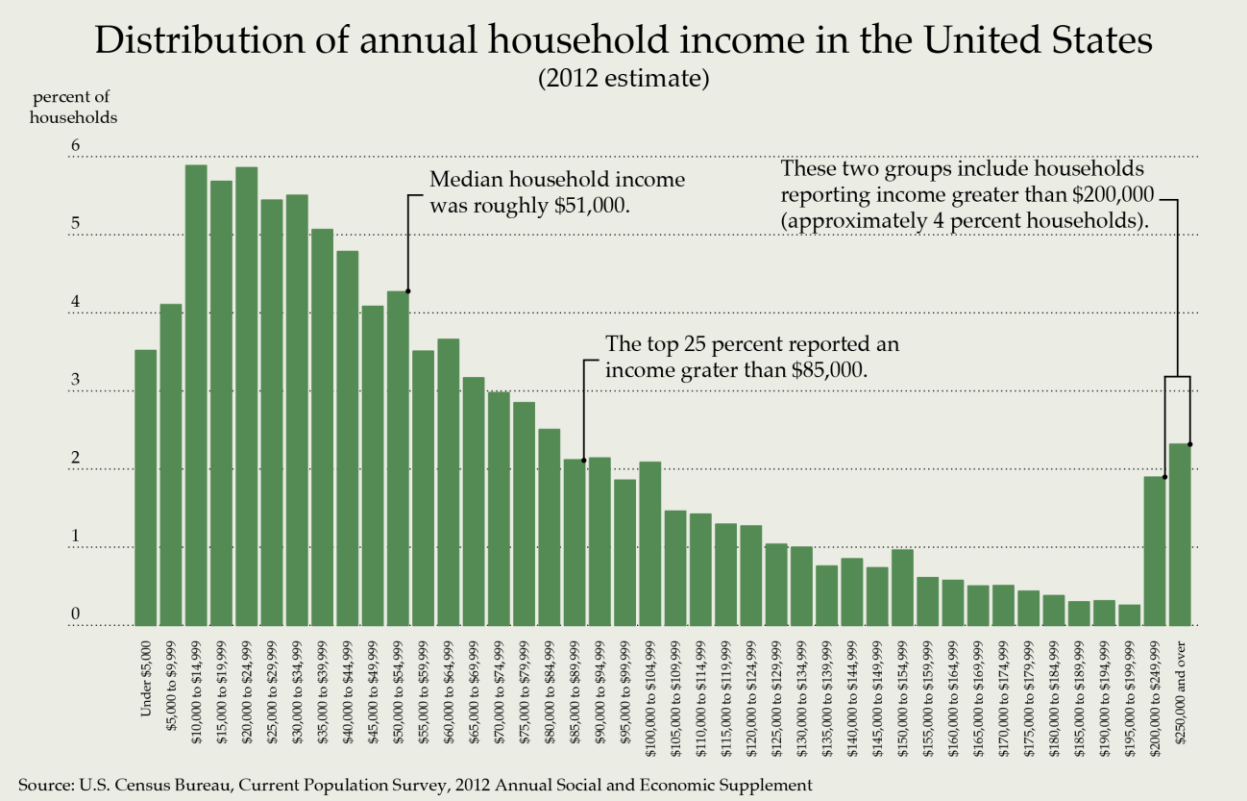
When the distribution is **negatively skewed** such that the tail of the distribution is to the left and the hump is to the right, the mean is less than the median which is less than the mode. The mode is always the value at the hump of the distribution; hence, in terms of the three measures of central tendency it will have the largest value. The mean is smallest, because the few extremely small scores will pull the mean down. The median is always the value at the 50th percentile.



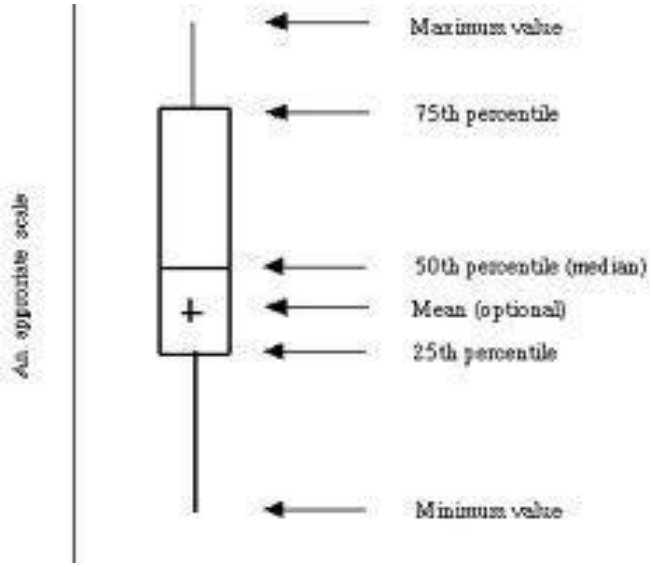
When a distribution is **positively skewed,** the tail of the distribution is to the right and the hump is to the left, because the mean is greater than the median which is greater than the mode. Again, the mode is always the value at the hump; hence, in terms of the three measures of central tendency it will have the smallest value. The mean will have the largest value, because the few extremely large scores will pull the mean up. The median is always the value at the 50th percentile; hence, it will always lie between the median and the mode.



Skew has to do with the pull of the individual values on the mean, so if the mean differs from the median/mode, the distribution is skewed. Because the median is always the 50th percentile and splits the distribution in half: the median is better to use as the measure of central tendency when the distribution is severely positively skewed or severely negatively skewed. An example of when the median is used as a measure of central tendency over the median is when reporting national incomes. For example, the graph below is a distribution of annual 2012 household incomes in the US, which shows substantial positive skew fewer high household incomes and most household incomes being relatively low. In this example, because of the high incomes the mean US household income would be about $68,000, but this mean is heavily influenced by the skew. The median, however, is about $51,000, which is much closer to the majority of household incomes. Hence, the median is a better measure of central tendency.



4.9 Box Plots

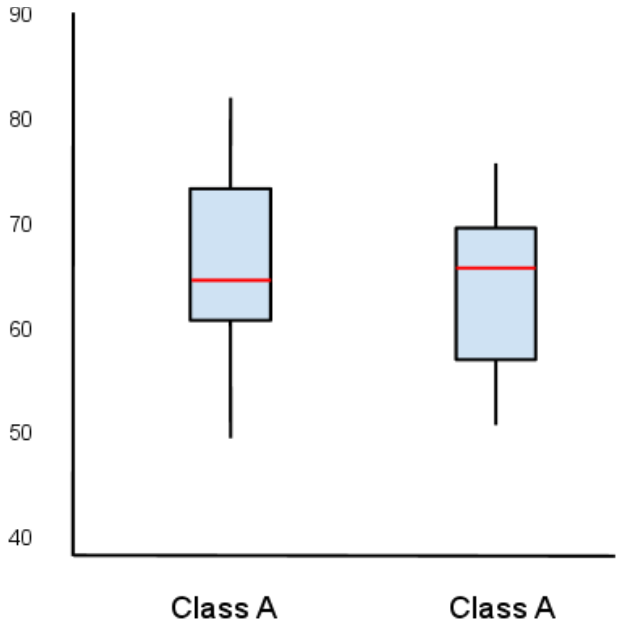


A **box plot** (***box-and-whisker plot***) is a way to present the dispersion of scores in a distribution by using five pieces of statistical information: the minimum value, the first quartile, the median, the third quartile, and the maximum value. Thus a box plot displays the location of the inter-quartile range ($Q_3 - Q_1$), the highest and lowest values, and the median. Box plots can also include the mean. The figure to the right is an example of the format of a box plot. (The 'appropriate scale' on the y-axis is simply the dependent variable.) In the graph, the rectangular box represents the location inter-quartile range within the overall range of values recorded on the dependent variable. The lower edge of the box represents the 25th percentile (Q_1) and the upper edge represents the 75th percentile (Q_3). The horizontal line within the box represents the median and the plus sign (+) represents the location of the mean, which is optional. The endpoints of the two lines that extend from the box represent where the minimum and maximum values occur in the distribution.

Let's say that we have two classes who took the same final exam, and I want to compare the dispersion of the scores in each class with box plots. The relevant statistics for each class are:

Statistic	Class A	Class B
Minimum Value	49	50
Q_1	60	58
Median	64	65
Q_3	73	69
Maximum Value	82	75

The box plots are below, presented in the same graph. The red lines represent the median:



The box plots show that the size of the inter-quartile range is about the same in each class, but it is shifted down slightly in class B. The medians are about the same for each class, and the overall range of scores is slightly less in Class B then in Class A. Thus, the box plots are a nice way to quickly examine the dispersion of the scores in a distribution. If the median line is equidistant from the edges of the box (i.e., equidistant from Q_1 and Q_3), then the data is not likely skewed. But, as can be seen in the box plots above, the median line is not equidistant from the edges, which suggests that the data are skewed. If the median line is closer to the bottom edge of the box (closer to Q_1), this suggests positive skew, as can be seen in Class A. If the median line is closer to the upper edge of the box (closer to Q_3), this suggests negative skew, as can be seen in Class B. Here's a little more on box plots: http://www.netmba.com/statistics/plot/box/

CH 4 Homework Questions

1. Identify and define the three measures of central tendency in your own words.

2. When will the mode most likely be used as a measure of central tendency?

3. Compute the median for the following scores: 7, 5, 10, 5, 5

4. Compute the median for the following scores: 4, 1, 6, 2, 11, 4

5. What does a measure of central tendency indicate regarding a distribution?

6. What does it mean that the mean is the mathematical center of a distribution?

7. Use the following data from eight individuals who were measured on variables X and Y, to calculate the requested sums, below.

i	X	Y
A	2	4
B	2	1
C	5	3
D	6	7
E	1	2
F	3	4
G	1	2
H	1	7

a. ΣX b. $\Sigma(X - Y)$ c. ΣX^2 d. $\Sigma(X)^2$
e. $(\Sigma X)^2$ f. $\Sigma(Y - X)^3$ g. ΣXY h. $\Sigma X - \Sigma Y$

8. Use the following data from eight individuals who were measured on variables X and Y, to calculate the requested sums, below.

i	X	Y
A	5	3
B	2	9
C	3	8
D	5	7
E	7	6
F	8	5
G	8	4
H	2	3

a. ΣX b. ΣY c. ΣXY d. ΣX^2
e. $(\Sigma X)^2$ f. $(\Sigma XY)^2$ g. $(\Sigma X)(\Sigma Y)$ h. $\Sigma(X - 3)$
i. $\Sigma(X - 2)(Y - 3)$ j. $\Sigma(Y - 3)^2$

9. Compute the mode, median, and mean for the following scores { 0, -1, 2, 1, -2, 0, -1, 0, -1, 0, 1, -1, 1, -2, 0, -1, 2, -1, 0, -1 }

10. Students were asked to rate their general political attitude on a scale from 1 (Extremely Liberal) to 11 (Extremely Conservative). The ratings from those students are displayed below.

1 2 2 2 3 3 3 3 3 3 4 5 5 5 5 5 6 6 6 6 6 7 8 8 8 9 9 9 9 9 9 9 10 11

a. What is the mode?
b. What is (are) the median's location(s)?
c. What is the median's value?
d. What is the location of the first quartile (Q_1)?
e. What is the location of the third quartile (Q_3)?
f. What is the value of the first quartile?
g. What is the value of the third quartile?
h. What is the inter-quartile range?

i. What is the sum of the scores equal to?
j. What is the sample mean equal to?
k. Is this distribution skewed, and if so, in what direction?

11. Construct a box plot for the data in #10. Include a '+' where the mean would be located.

12. *Use the data below to answer the questions that follow*. Here are the numbers of wins the New York Yankees for each of the last 10 seasons:

Year	Wins
2012	95
2011	97
2010	95
2009	103
2008	89
2007	94
2006	97
2005	95
2004	101
2003	**101**

a. Determine the mode.
b. Determine the median value.
c. Calculate the sample mean.
d. Is the sample of wins positively skewed, negatively skewed, or unskewed? *How do you know*?

13. *Use the data below to answer the questions that follow*. Here are the numbers of wins the New York Yankees for each of the 10 seasons that preceded those in #12.

Year	Wins
2002	103
2001	95
2000	87
1999	98
1998	114
1997	96
1996	92
1995	79
1993	88
1992	76

a. Determine the mode.
b. Determine the median value.
c. Calculate the sample mean.
d. Is the sample of wins positively skewed, negatively skewed, or unskewed? *How do you know*?

14. Compute the mean for the following five scores: 10, 11, 12, 13, 14. Now, generate a new set of five scores by adding a constant of 3 to each original score. Compute the mean for the new scores. Do the same for another set of five scores generated by subtracting a constant of 10 from each original score. Do the same for another set of five scores generated by multiplying each original score by a constant of 2. What are the effects of adding a constant to each score, subtracting a constant from each score, and multiplying each score by a constant in a set of scores?

15. *Use the data below to answer the questions that follow*.

8 9 7 10 6 9 8 9 5 9

a. Determine the mode.
b. Determine the median value.

c. Calculate the sample mean
d. Is the sample of wins positively skewed, negatively skewed, or unskewed? *How do you know*?

16. Using the data in #15, add a constant of 5 to each value and then answer each of the following questions.

a. Determine the mode.
b. Determine the median value.
c. Calculate the sample mean.
d. Is the sample of wins positively skewed, negatively skewed, or unskewed? *How do you know*?

17. Suppose you measured the mean time it took ten children to solve a problem and found it to be 45.28 seconds. You later discovered, however, that your timing device (a watch) was 3 seconds too slow for each child. What was the real mean score?

18. Compute the mean for the following five scores: 20, 50, 30, 10, 40. Now generate a new set of five scores by multiplying each original score by 3. Compute the mean for the new scores. Do the same for another set of five scores generated by dividing each original score by a constant of 10. What are the effects on the mean of multiplying or dividing each score in a set of scored by a constant?

19. In your own words, what are the advantages and disadvantages of using the mean a measure of central tendency? The median? The mode?

20. Which measure of central tendency is influenced by extreme scores (outliers)? Which is not influences by outliers?

21. Take a look at these two sets of scores. In which case is the mean a poorer measure of central tendency? *Why*?

Set A			Set B	
65	65		78	79
66	200		75	78
64	65		76	76
68	65		80	77
70	64		82	75

22. Given a set of scores where the mode is 9, the median is 15 and the mean is 21, are these scores skewed? If so, how?

23. If the mode was 23, the mean was 20 and the median was 22, would these scored be skewed If so, how?

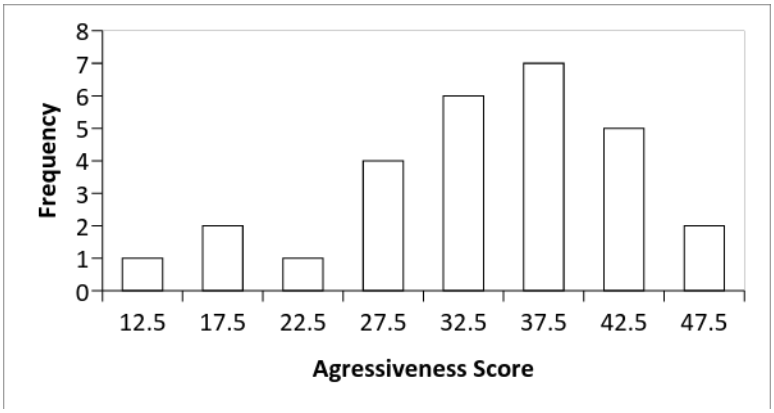
24. If the mode was 20, the median was 20, and the mean was 20, would the scores be skewed? If so how?

25. A teacher wants to change seating arrangement of students in her class in the hope it will increase the number of comments made by students. She decides to see how many comments students make with the current seating arrangement. A record of the number of comments made by 8 of her students during one class period is shown below. She wants to summarize this data by computing the typical (mean) number of comments made that day. Explain how reporting the mean number of comments could be misleading.

	Student Initials							
Number of	AA	RF	AG	JG	CK	NK	JL	AW
Comments	0	5	2	22	3	2	1	2

26. Explain how it could happen that if one person moves from city A to city B, it is possible for the average (mean) IQ in both cities to increase.

27. A test to measure aggressive tendencies was given to a group of teenage boys who were members of a street gang. The test is scored from 10 to 60, with a high score indicating more aggression. The histogram represents the results for these 28 boys. Which do you think will be larger, the mean or the median score? Or do you think they will be similar? Explain.



Chapter 5: Variability

5.1 What is Variability? Why is it Measured?

Look at the picture above. The variety of colors, shapes, sizes, flavors, and spiciness of those chili peppers is beautiful! **Variability** is differences among items, which could be differences in eye color, hair color, height, weight, sex, intelligence, etc. There is variability in the number of hairs on your head across the days of the week as we gain some and we lose some, and there is variability in how you feel across the day. Variability is the second main important topic in statistics; central tendency being the other. If central tendency (mean, median, mode) estimate where a distribution falls, variability measures the dispersion and the similarity among scores in a distribution.

Why do we measure variability? Recall, the mean takes into account all scores in a distribution. But even if you know where the mean falls you do not know anything about the actual scores. That is, if I told you the mean of a set of n = 10 scores on a quiz with a range of 0 – 10 was M = 7, but told you nothing else, could you tell me how alike those scores were? The No, of course not. Even though the mean accounts for all scores in a distribution, knowing the mean does not tell you anything about the actual individual scores; the mean just tells you were a distribution of data tends to fall. For example, if the mean for a set of ten scores on a statistics quiz is M= 7, those ten scores could be:

7 + 7 + 7 + 7 + 7 + 7 + 7 + 7 + 7 + 7 = 70/10 = 7, or

6 + 6 + 6 + 6 + 6 + 8 + 8 + 8 + 8 + 8 = 70/10 = 7, or

1 + 1 + 5 + 5 + 9 + 9 + 10 + 10+ 10+ 10 = 70/10 = 7,

...or any other random combination where the n = 10 scores add to 70 (assuming the range is 0 – 10). Even though the mean takes individual scores into account, it does not measure the variability.

Statistically speaking, variability is how alike scores are in a set of data. If a measure of variability is small (approaching zero), it indicates low variability and most scores are alike; whereas a large measure of variability (approaching infinity) indicates scores that are different. There are several measures of variability. In the following sections I start with the least complex measure and move to the more complex measures. Each successive measure has all the characteristics as the less-complex measures plus something else. Measures of variability include the **range**, **sum of squares**, the **variance**, and the **standard deviation**.

5.2 Range

In each of the following sections, the measures of variability are calculated for both of the following sets of n = 10 scores. Note that in both sets the sum of scores and the mean are identical:

Set I:	2	2	2	4	5	5	6	6	8	10	$\sum X = 50$	$M = 5$
Set II:	4	4	5	5	5	5	5	5	6	6	$\sum X = 50$	$M = 5$

The **range** is the largest value minus the smallest value. For Set I the range is 10 – 2 = 8, and the range for Set II is 6 – 4 = 2. The range provides information only about the overall dispersion of scores, that is, the area a distribution covers. A large range indicates a large spread of scores compared to a small range; however, the range does not say anything about individual values. For example, the following sets of n = 5 scores have the same range (range = 9), but different underlying distributions:

Set A: 1 1 1 1 10 Set B: 1 2 4 5 10

Clearly there are more differences among scores in Set B than Set A, but the range does not account for this variability. What we need is some way to measure the variability among the scores.

5.3 Sum of Squares

The **sum of squares** (**SS**) is defined as the *sum of the squared deviation scores from the mean* and measures the summed (total) variation in a set of data. The formula for the sum of squares is:

$$SS = \sum (X - \bar{X})^2$$

Sum of squares could be equal to zero if there is no variability and all the scores are equal; but, the sum of squares can never be negative. To calculate the sum of squares, first subtract the mean from each score in the data, a process called **mean centering**. After mean centering each score, square each deviation from the mean, and then sum all the squared deviations. This is done for Sets I and II below:

Set 1		
X	$(X - \bar{X})$	$(X - \bar{X})^2$
10	5	25
8	3	9
6	1	1
6	1	1
5	0	0
5	0	0
4	-1	1
2	-3	9
2	-3	9
2	-3	9
		SS = 64

Set 2		
X	$(X - \bar{X})$	$(X - \bar{X})^2$
6	1	1
6	1	1
5	0	0
5	0	0
5	0	0
5	0	0
5	0	0
5	0	0
4	-1	1
4	-1	1
		SS = 4

The sum of squares for Set 1 is SS = 64 and for Set 2 is SS = 4. Larger sum of squares indicates there is more variability among the scores in Set 1. Sum of squares measures the total variation among scores in a distribution; thus, sum of squares does not measure average variability, which is what we want. We want a measure of variability that takes into account both the variation of the scores and number of scores in a distribution; hence, a measure of 'average variability' of the scores in a distribution, which is the **variance**.

5.4 Variance (s^2)

Variance (s^2) is defined as the *average sum of the squared deviation scores from a mean*, or samply the *average sum of squares* and measures the average variability among scores in a distribution. Thus, variance measures by how much scores in a distribution differ on average. The formula for sample variance is:

$$s^2 = \frac{\Sigma(X - \overline{X})^2}{n}$$

To calculate sample variance, first, calculate the sum of squares as in Section 5.3. Next divide the sum of squares by the number of scores in the data set (n); hence, once you calculate the sum of squares finding the sample variance is just dividing SS by n:

$$s^2 = \frac{SS}{n}$$

From the data in Section 5.3 we have, Set 1: SS = 64, *n* = 10; and Set 2: SS = 4, *n* = 10. The variance for each data set is:

$$Set1: \quad s^2 = \frac{64}{10} = 6.4 \qquad\qquad Set2: \quad s^2 = \frac{4}{10} = 0.4$$

Each value measures the average variability among the scores in each set of data, where larger measurements of variance indicate greater variability. Indeed, you can see this by comparing Sets 1 and 2; there is more variability and larger deviations between the scores in Set 1 compared to Set 2

One issue is variance is in squared units, not the original unit of measurement. Anytime you square a value you change the original unit of measure into a squared measurement units. This is more easily understood if you think of the scores in Set I as measurements of length in centimeters (cm). If the original measurements in Sets 1 and 2 were of length in centimeters, when the values were squared the measurement units are now cm^2, which represents measurements of area (see table to right)

X	$(X - \overline{X})$	$(X - \overline{X})^2$
10 cm	5 cm	25 cm^2
8 cm	3 cm	9 cm^2
6 cm	1 cm	1 cm^2
6 cm	1 cm	1 cm^2
5 cm	0 cm	0 cm^2
5 cm	0 cm	0 cm^2
4 cm	-1 cm	1 cm^2
2 cm	-3 cm	9 cm^2
2 cm	-3 cm	9 cm^2
2 cm	-3 cm	9 cm^2

What is the solution to the issue of squaring both the value and the unit of measure; how do we get the measurement units back from a squared unit into the original unit of measurement? If you guessed take the square root of the variance, you are correct. We call this the standard deviation.

5.5 Sample Standard Deviation (s)

Once sample variance has been calculated, calculating the **sample standard deviation** (*s*) is as simple as taking the square root of the variance; thus, the standard deviation is in the original unit of measure. The

sample standard deviation is defined as *the square root of the average sum of squared deviations from a mean*; or simply, *the square root of the variance*. The formula for the sample standard deviation is:

$$s = \sqrt{\frac{\Sigma(X-\overline{X})^2}{n}} \quad \text{or} \quad s = \sqrt{\frac{SS}{n}} \quad \text{or} \quad s = \sqrt{s^2}$$

In each formula, all you are doing is taking the square root of the variance. From the data in Sections 5.3 and 5.4, we have, Set 1: SS = 64, n = 10, s^2 = 6.4; and Set 2: SS = 4, n = 10, s^2 = 0.4. The sample standard deviations in Sets I and II are:

$$\square = \sqrt{6.4} = 2.53 \qquad \square = \sqrt{0.4} = 0.632$$

The standard deviation measures the average deviation (difference) between any score from the data and the mean of a distribution. For example, in Set 1, a randomly selected score is expected to deviate from the mean by 2.53 and a score is expected to deviate from the mean by 0.632 in Set 2. This does not mean you will find scores in the data set that deviate from the mean by exactly these amounts; the standard deviation is an average value. Because the standard deviation measures the variability of scores in the original unit of measure it is our best measure of variability. Note that the standard deviation takes into account all the scores in a distribution, because it is based on the variance and the sum of squares.

5.6 The Population Formulas for Variability

Calculating sum of squares in a population is no different than calculating the sample sum of squares, the only thing that differs is the symbols. The formula for sum of squares in a population is:

$$\square\square = \sum \quad (\square - \square)^2$$

This is functionally equivalent to the formula for the sample sum of squares in Section 5.3, the only difference is that here you subtract a population mean (μ) from each score. Thus, the only difference between these formulas is working with sample data versus population data. The formulas for the **population variance** (σ^2, 'little sigma squared') and the **population standard deviation** (σ, 'little-sigma') are functionally equivalent to the formulas for calculating the sample variance (s^2) and the sample standard deviation (s), respectively. The formulas for population variance are:

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N} \quad \text{or} \quad \sigma^2 = \frac{SS}{N}$$

And for population standard deviation are:

$$\sigma^2 = \sqrt{\frac{\Sigma(X-\mu)^2}{N}} \quad \text{or} \quad \sigma^2 = \sqrt{\frac{SS}{N}} \quad \text{or} \quad \sigma^2 = \sqrt{\sigma^2}$$

The important point is the concepts *sum of squares* (total variability), *variance* (average variability) variance), and *standard deviation* (average difference between a score and the mean) do not change when you are working with a sample or population (the same is true for measures of central tendency). Calculating a measure of variability or o central tendency in a sample is the same as calculating a measure of central tendency in a population and the measures mean the same thing, but refer to different types of groups. The mathematical procedures do not change, just the symbols. Too many times I have seen students get hung up on the symbols used in population and samples and think the different formulas measure different things; however, the concepts variance, standard deviation, mean, median, etc., and the methods for calculating them are the same in a sample and a population. The point is to not get caught up in the different symbols. Concentrate on what the concepts mean, the language of statistics, and what each measure reflects, and you'll be fine.

5.7 Scaling Scores

There may be times when you must change each score in a data set by adding, subtracting, multiplying, or dividing by a constant. As long as you perform the same operation to each score, this is **scaling** your scores. There are few interesting things that changing each score in the data set by a constant will do to the original mean and standard deviation. That is, if you change each score in a data set by a constant and in the same manner, you will not have to completely recalculate the mean and standard deviation.

If you add or subtract a constant to each score in a set of data, the mean of that scaled set of data will increase or decrease by that constant, respectively, from the original mean, while the standard deviation does not change. For example, if you add X = 2 to each score from Set I (see table below), the original mean of 5 will increases by 2 (new mean = 7). If you subtract 2 from each score the original mean decreases by 2 to 3. In both cases, the standard deviation does not change.

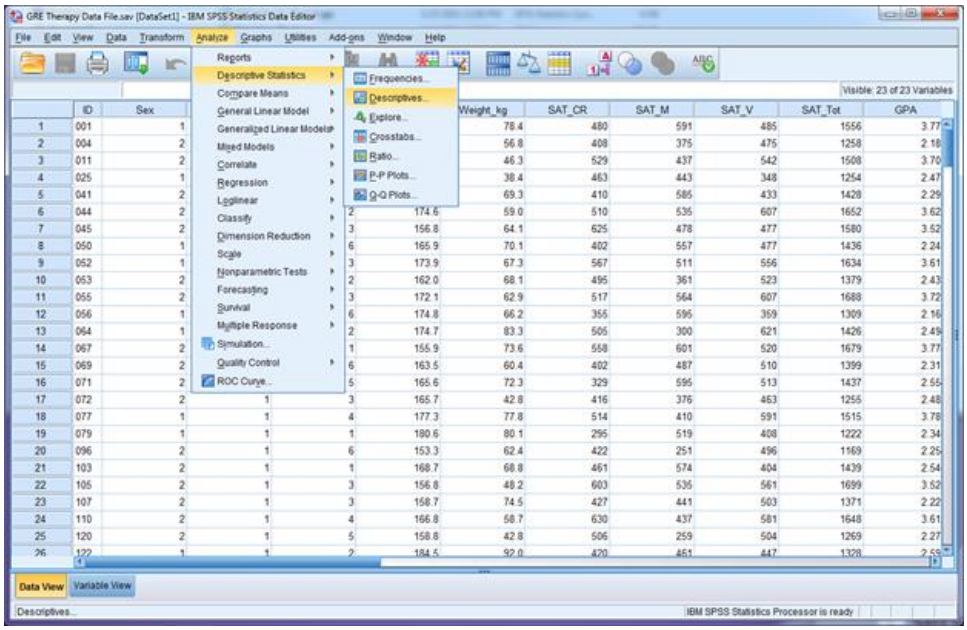
If you multiply each score by a constant, the new mean is equal to the original mean multiplied by that constant and the new standard deviation will be equal to the original standard deviation multiplied by that constant. For example if you multiply each score from Set I by 2, the original mean of 5 is now equal to 10 and the original standard deviation (2.53) increases to 5.06.

If you divide each score in a data set by a constant, the new mean will be equal to the original mean divided by that constant, and the new standard deviation will be equal to the original standard deviation divided by that constant. For example if you divide each score from Set I by 2, the original mean (5) is now equal to 2.5, and the original standard deviation (2.53) is now equal to 1.265. You can check all of this using the original data from Set I and the scaled Set I scores below::

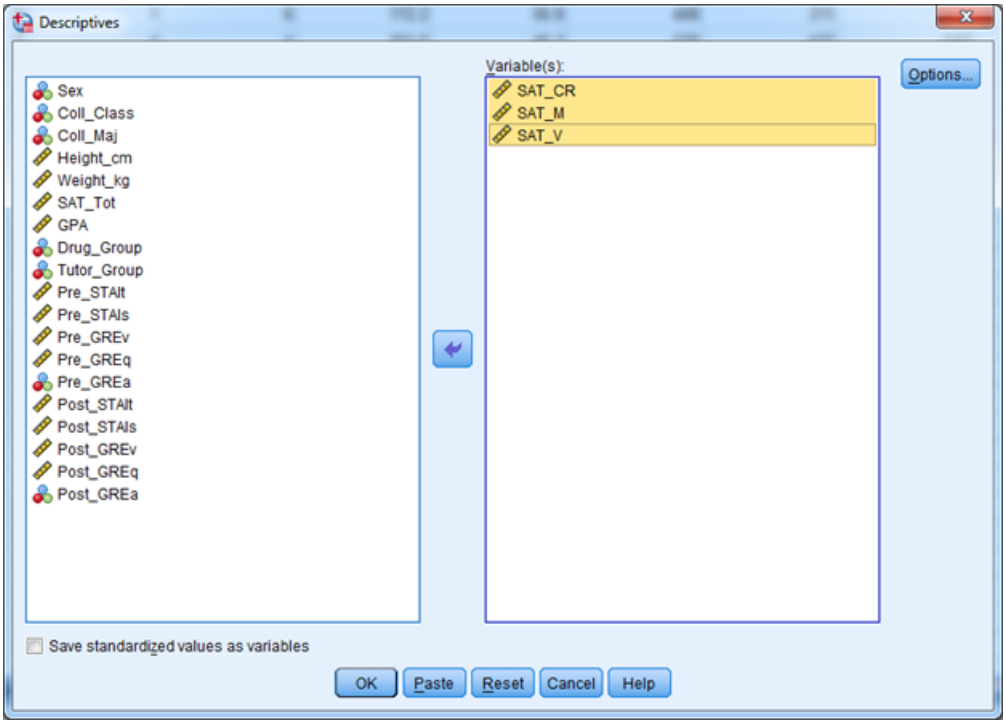
X	**X + 2**	**X - 2**	**X*2**	**X/2**
10	12	8	20	5
8	10	6	16	4
6	8	4	12	3
6	8	4	12	3
5	7	3	10	2.5
5	7	3	10	2.5
4	6	2	8	2
2	4	0	4	1
2	4	0	4	1
2	4	0	4	1
∑X = 50	∑X = 70	∑X = 30	∑X = 100	∑X = 25
M = 5	*M* = 7	*M* = 3	*M* = 10	*M* = 2.5
s^2 = 6.4	s^2 = 6.4	s^2 = 6.4	s^2 = 25.6	s^2 = 1.60
s = 2.530	s = 2.530	s = 2.530	s = 5.060	s = 1.265

5.7 Central Tendency and variability in SPSS

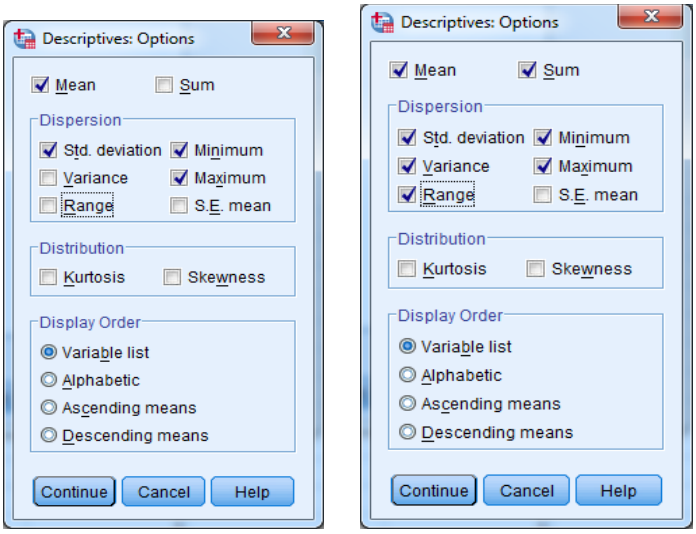
The following uses the GRE Therapy Data file. This data file is based on a hypothetical study examining the influences of a study-aid drug and types of tutoring on performance on the Graduate Record Examinations (GREs). To request SPSS calculate measures of central tendency and variability, from the Analyze menu, select Descriptive Statistics, and then Descriptives:



Say we want descriptive statistics on the SAT Critical Reading (SAT)CR), SAT Math (SAT_M), and SAT Verba; (SAT_V) test scores . Move those three variables from the left to the blank area on the right:



Next, click the Options button. By default SPSS calculates the mean and standard deviation, and provides the minimum and maximum values; however, you can have SPSS calculate other descriptive statistics by checking appropriate boxes. Go ahead and check the boxes for the Sum, Variance, and Range (S.E. Mean is covered in chapter 7). Click continue, and in the main window click OK:



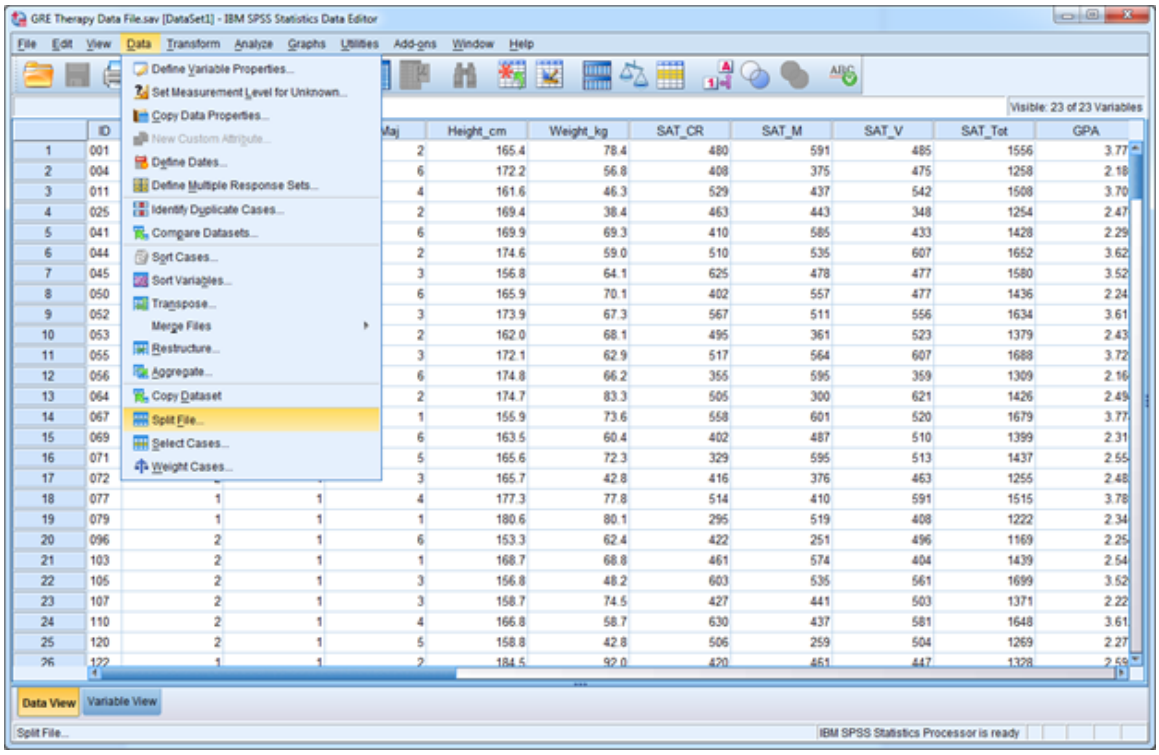
The outcome is below, where it can be seen that 240 students contributed to each variable, and while the mean critical reading (*M* = 491.14) and verbal scores (*M* = 496.69) are similar, they are slightly less than the mean math score (*M* = 516.79):

Descriptives

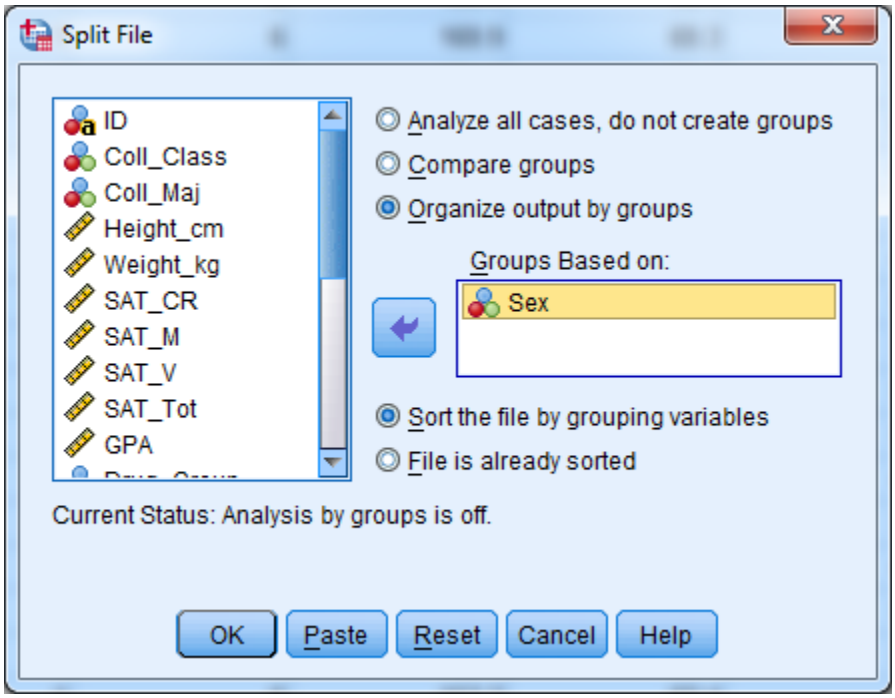
Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean	Std. Deviation	Variance
SAT_CR	240	624	237	861	117874	491.14	105.938	11222.800
SAT_M	240	757	191	948	124029	516.79	127.100	16154.511
SAT_V	240	291	348	639	119205	496.69	66.772	4458.458
Valid N (listwise)	240							

What if you wanted to examine performance for males and females separately for the same variables ? In this case, we need to From the Data menu select Split File:



Click the button for Organize output by groups, and then move the variable Sex into the blank area under Groups Based on (see right). This tells SPSS to analyze the data separately for any groups defined within the variable Sex. Click the OK button, and then go back through the procedure above for requesting descriptive statistics.



The output of this analysis is below, where you can see separate sets of output for Female students and male students:

Descriptives

Sex = Males

Descriptive Statistics[a]

	N	Range	Minimum	Maximum	Sum	Mean	Std. Deviation	Variance
SAT_CR	109	556	237	793	53807	493.64	102.428	10491.491
SAT_M	109	739	209	948	58916	540.51	142.506	20307.863
SAT_V	109	280	348	628	53779	493.39	72.780	5296.887
Valid N (listwise)	109							

a. Sex = Males

Sex = Females

Descriptive Statistics[a]

	N	Range	Minimum	Maximum	Sum	Mean	Std. Deviation	Variance
SAT_CR	131	589	272	861	64067	489.06	109.120	11907.073
SAT_M	131	611	191	802	65113	497.05	109.378	11963.567
SAT_V	131	279	360	639	65426	499.44	61.477	3779.463
Valid N (listwise)	131							

a. Sex = Females

CH 5 Homework Questions

1. Identify and define the three main measures of variability:

2. What is the problem with using the sum of squares as a measure of variability, that is, why is variance a better measure?

3. What is the standard deviation our best measure of variability?

4. If the variance of a set of scores is 10000.00 what is the standard deviation?

5. If the standard deviation of a set of scores is 9.00 what is the variance?

6. Without calculating it, what must the variance of the following scores equal: 4, 4, 4, 4? What must the standard deviation equal? What must the sum of squares equal? Why?

7. The following scores were obtained on a quiz: { 3 5 5 6 7 7 7 8 8 9 }
   a. Calculate the mean of the scores.
   b. Calculate the sum of squares.
   c. Calculate the sample variance.
   d. Calculate the standard deviation.

8. The following scores were obtained on another quiz { 3 4 4 6 7 7 8 8 9 9 }
   a. Calculate the mean of the scores.
   b. Calculate the sum of squares.
   c. Calculate the sample variance.
   d. Calculate the standard deviation.

9. For the sets of data in exercises 7 and 8, what descriptive statistics are the same (if any) and which are different (if any)? What does this tell you about the measures of central tendency and the measures of variance?

10. *Use the data below to answer the questions that follow*. Here are the numbers of wins the New York Yankees for each of the last 10 seasons:

Year	**Wins**
2012	95
2011	97
2010	95
2009	103
2008	89
2007	94
2006	97
2005	95
2004	101
2003	**101**

a. Calculate the sum of squares.
b. Calculate the sample variance.
c. Calculate the sample standard deviation
d. Calculate the estimate of the population variance.
e. Calculate the estimate of the population standard deviation.

11. *Use the data below to answer the questions that follow*. Here are the numbers of wins the New York Yankees for each of the 10 seasons that preceded those in #10:

Year	**Wins**
2002	103
2001	95
2000	87

1999	98
1998	114
1997	96
1996	92
1995	79
1993	88
1992	76

a. Calculate the sum of squares.
b. Calculate the sample variance.
c. Calculate the sample standard deviation.
d. Calculate the estimate of the population variance.
e. Calculate the estimate of the population standard deviation.

12. *Use the data below to answer the questions that follow*.

8 9 7 10 6 9 8 9 5 9

a. Calculate the sum of squares.
b. Calculate the sample variance.
c. Calculate the sample standard deviation.

13. Using the data in #12, add a constant of 5 to each value and then answer each of the following questions.

a. Calculate the sum of squares.
b. Calculate the sample variance.
c. Calculate the sample standard deviation.

14. Using the data in #12, multiple each value by a constant of 2 and then answer each of the following questions.

a. Calculate the sample mean.
b. Calculate the sum of squares.
c. Calculate the sample variance.
d. Calculate the sample standard deviation.

15. Compute the variance and standard deviation for the following eight scores: { 6, 10, 8, 8, 10, 8, 6, 8 }. Now, generate a new set of five scores by adding a constant of 3 to each original score. Compute the variance and standard deviation for the new scores. What are the effects of adding a constant on the variance and on the standard deviation?

16. How does the standard deviation help you interpret a set of data?

Use the following to answer #17 – 20: A political scientist studied how satisfied people were with the President's job performance and Congress' job performance. Each person was given a job performance satisfaction test on which scores ranged from 1 to 7, with higher scores indicating more satisfaction. The scores were:

Individual	*President's Job Performance*	*Congress Job Performance*
A	2	1
B	5	3
C	6	3
D	7	5

E	5	3

17. Compute the mean of the satisfaction ratings for the President's and Congress' job performance.

18. Compute the standard deviation for the satisfaction ratings for the President's and Congress' job performance.

Individual	*President's Job Performance*			*Congress Job Performance*		
	X	(X – M)	$(X - M)^2$	X	(X – M)	$(X - M)^2$
A	2	-3	9	1	-2	4
B	5	0	0	3	0	0
C	6	1	1	3	0	0
D	7	2	4	5	2	4
E	5	0	0	3	0	0

19. Based on the results, whose job performance are people more satisfied with? Why?

20. Based on the results, which job performance ratings are more consistent? Why?

21. Suppose you measured the weights of 100 people and found a mean of 150 with standard deviation 11. Then you learned that your scale was 2 pounds to heavy. What were the correct mean and standard deviation?

22. How accurate are people ate estimating length? In an experiment, people are presented with a length that was 25-cm in length for five seconds and asked to estimate the length of the rope. Ten people gave the following estimates (in cm):

16 17 33 31 25 22 30 34 20 32

Compute the mean and standard deviation for these data. How accurate are the estimated considering the mean score across all participants? How does the standard deviation help to interpret the mean?

23. Following up on #22, in a second experiment, people are presented with the same 25-cm piece of rope, but for one minute before making their estimates of the ropes length. Ten people gave the following estimates (in cm):

23 27 25 23 26 24 26 22 23 26

Compute the mean and standard deviation for these data. How accurate are the estimated considering the mean score across all participants? How does the standard deviation help to interpret the mean?

24. An organizational psychologist studied how satisfied employees were in two different companies. All employees were given a job satisfaction test on which scores could range from 1 to 7, with higher scores indicating greater satisfaction. The scores were as follows:

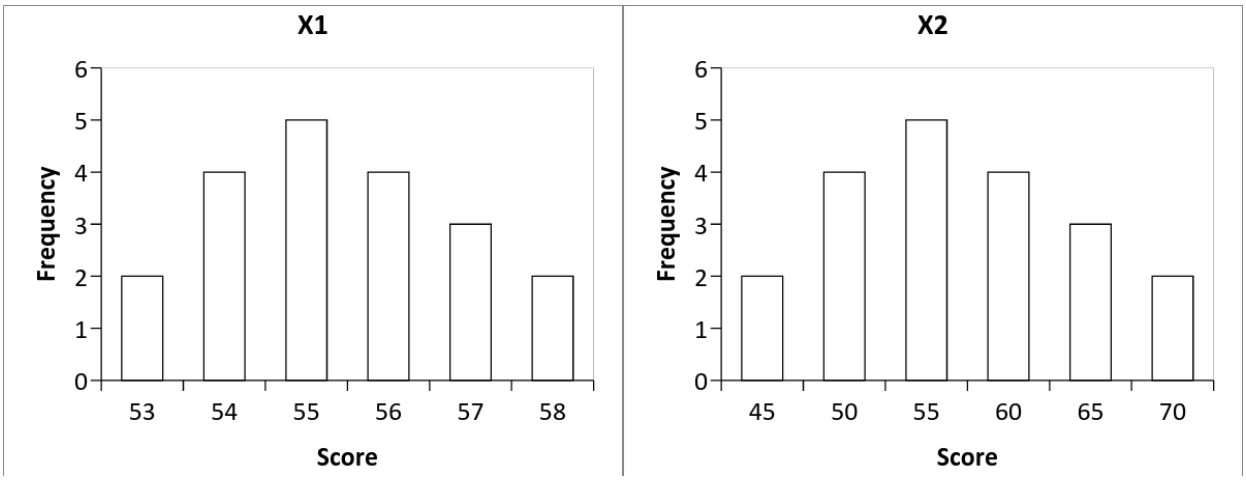
Company A:	4	6	5	4	3	2
Company B	4	4	4	4	4	4

Compute the mean and standard deviation for each company. Based on the results, compare employee satisfaction in the two companies. How do the standard deviations help to interpret the means?

25. Suppose scores on a math achievement test approximate a normal distribution with *M* = 500, min = 350, and max = 650. Estimate the standard deviation of this distribution.

26. Two students who took a statistics class received the following Quiz scores (each Quiz is out of 100 points): Student A: 60, 90, 80, 60, 80 and Student B: 40, 100, 100, 40, 90. If you had an upcoming statistics test, who would you have rather had as a study partner, A or B? Support your answer.

27. Consider the two histograms displayed below. The histogram labeled X1 has a mean of 54 and the histogram labeled X2 has a mean of 53. Please indicate which one has a larger standard deviation and WHY that histogram has the larger standard deviation. (It is not necessary to do any calculations.)



Chapter 6: Standard Scores and Normal Distribution

"When you told me on the phone you were 42, 22, 38 I didn't realize you meant your age, your I.Q., and your shoe size."

6.1 Comparing Scores across Distributions

The picture above does a great job of introducing the concept of **standard scores** and why comparing across distributions is tricky. The gentleman was under the impression the numbers the lady gave him meant something other than what she meant, that is, how the gentleman interpreted 42 was not what the lady meant as 42. Thus, the cartoon illustrates that when comparing across distributions, you cannot rely on raw scores; rather, raw data must be converted into a standard measure that will be understandable in any distribution.

To understand why comparing across distributions is problematic, consider the following situation: You have a friend, Stu, who thinks he is smarter than you. Both you and Stu take the same 19th Century American Literature course from the same professor, in the same semester, but in different class sections. You have a test in that course on the same day and the content of the test is exactly the same. On the test, Stu obtains 85 out of 100 points and you obtain 80 out of 100 points. Of course, Stu starts bragging and boasting about his 19th Century American Literature intelligence. But who *really* did better on this test?

It's true your score of 80 was less than Stu's score of 85, but what happens when you consider information about each class' test score distribution? Say each test was on the same material and included the same content, but the makeup of each test was slightly different (e.g., more multiple choice questions for one section, slightly differently-worded questions). Indeed, maybe the classes were a little different since there are different students in each section and the professor may adjust change lessons accordingly. That said, assume your class' mean test grade was $M = 75$ points and Stu's class mean was $M = 80$ points. Now who did better? When you take the mean of each class into account, it seems you and Stu performed equally, because both scores were 5 points above your respective class mean. However, because the means of each class are different, this suggests the underlying test score distributions in each class were also different; hence, a 5 point difference in one class may be different than a 5 point difference in the other. Thus, the 5 point differences must be considered *relative* to the class distribution.

So, who did better, you or Stu? What you need is a 'standard measure' to compare your scores across the two class distributions. Fortunately, we already know of such a measure: the *standard deviation*. Remember, the standard deviation is the average difference between a score (x) and the mean (M) of a distribution. Although the value of the standard deviation changes depending on the data in the distribution, a standard deviation means the same thing in any distribution. For example, one standard deviation from the mean means the same thing regardless of the underlying distribution. Similarly, knowing a score is two standard deviations away from the mean would mean the same thing in any distribution. Thus, the standard

deviation measures the distance between a score and the mean of a distribution and that distance can be used to compare across different distributions, because standard deviations have a universal meaning.

When you measure the number of standard deviations between your test score and the mean of your class, you can compare that difference, which is in numbers of standard deviations, to the number of standard deviations between Stu's test score and the mean of his class. The question is how do you do this? The measure of the number of standard deviations between any raw score and the mean of a distribution is a **standard score**, or **z-Score**.

6.2 Standard (z) Scores

A **standard score** (**z-Score**) measures the number of standard deviations between a score (X) and the mean (*M* or μ). To calculate a standard (z) score for a score, you need the mean of the distribution (*M* or μ), the standard deviation (*s* or σ), and a raw score (X). The formula for calculating a standard score is:

$$z = \frac{X - \mu}{\sigma}$$

And in a sample is:

$$z = \frac{X - \overline{X}}{s}$$

Both formulas are functionally equivalent: subtract the mean from a score (X) and divide that difference by the standard deviation. The 'z' comes from a standard scores when it is calculated from a raw score in a normal distribution (more on this later).

Back to you and Stu: Assume the standard deviation for your class is *s* = 1 and the standard deviation for Stu's class is *s* = 5. Remember, your score was X = 80 and Stu's was X = 85. Your class' mean was *M* = 75 and Stu's class mean was *M* = 80. We have:

$$z_{You} = \frac{80 - 75}{1} = 5 \qquad z_{Stu} = \frac{85 - 80}{5} = 1$$

The difference between each raw score and the mean of each class is identical (difference = 5), but Stu's standard score is less than yours (*z* = 1 vs. *z* = 5). The reason has to do with the difference in the variability among the test scores between classes. The standard deviation was smaller in your class than Stu's class, and a smaller standard deviation suggests less variability among scores in your class compared to Stu's. Less variability indicates the scores in your class were, on average, closer to the mean, compared to Stu's class where the higher variability indicates greater differences from the mean. Stated differently, the higher standard deviation in Stu's class suggests a greater dispersion of scores around the mean. Because most scores in your class were similar to the mean, deviations from your class' mean are unlikely, but because there is higher variability in Stu's class deviations from the mean are not surprising.

Standard scores can be used to compare test performance across the two class distributions. Because your score is farther away from the mean than Stu's score, in standard deviations you did better on the test. Thus standard scores provide a way to compare across distributions. In short, because raw scores mean different things in different distributions, it is appropriate to convert scores (X) into standard (z) scores when comparing across distributions.

6.3 Characteristics of Standard (z) Scores

There are several noteworthy characteristics of standard scores. First, standard scores can be used to compare across distributions as was done in the example in Section 6.2. Again, when comparing across

distributions, unless the mean and standard deviation are identical in each distribution or if the raw scores come from the same distribution, it is inappropriate to compare scores; you should use standard scores.

Second, larger standard scores indicate larger deviations (distances) from the mean. Thus, the larger the standard score, the farther the raw score (X) is from the mean. And, of course, a standard score of zero (*z* = 0) indicates a score lies at the mean.

Third, standard scores are defined as the *distance between a score and its mean in standard deviations*; thus, your z-score from Section 6.2 of z = 5 indicates your score of X = 80 is five standard deviations above your class' mean, and Stu's z-score of z = 1 indicates his score of X = 85 is one standard deviation above his class' mean. Similarly, a z-Score of z = -1.5 indicates a score is one and one-half standard deviations *below* the mean.

Fourth, standard scores can be positive or negative. Positive z-scores indicate the score was greater than the mean and negative z-scores indicate the score was less than the mean. Importantly, the sign (+/-) does not say anything about magnitude of z-score. For example, a z-score of -2.0 is larger than a z-score of +1.0, because, z = -2.0 is twice the distance from the mean compared to z = +1.0. Thus, the sign (+/-) of a z-score indicates the direction a score from the mean; the absolute value indicates magnitude.

Finally, if you calculate the mean, variance and standard deviation of a distribution of z-scores calculated from a set of scores, the mean of the z-distribution will be zero and the variance and standard deviation will be equal to one:

X	$(X - \bar{X})$	$(X - \bar{X})^2$	**z**	$(z - \bar{z})$	$(z - \bar{z})^2$
4	1	1	1.118	1.118	1.25
4	1	1	1.118	1.118	1.25
3	0	0	0	0.000	0
2	-1	1	-1.118	-1.118	1.25
2	-1	1	-1.118	-1.118	1.25
$\sum X=15$		SS = 4	$\sum z = 0$		$SS_z = 5$
M = 3		$s^2 = 0.8$	$\bar{z} = 0$		$s_z^2 = 1$
		s = 0.894			$s_z = 1$

6.4 Using the Mean, Standard Deviation, and Z to find Raw Scores

You can use the standard deviation to determine the distance between a score and the mean, but what if you want to determine the score that would be a certain number of standard deviations above or below a mean? That is, say we have a distribution of 19th Century American Literature test scores with mean *M* = 70 and standard deviation *s* = 10. Assume that any score 2.5 standard deviations or more above the mean earns an A; that is, the minimum score for an A is associated with z = 2.5. So, we want to know the score (X) that is exactly 2.5 standard deviations above the mean. How can we do this?

Remember a z-score measures the number of standard deviations a raw score is away from a mean. In the example above we are looking for the raw score associated with z = 2.5. To calculate this raw score we use one of the two following equations, both of which are used to determine the raw score associated with a z-score, given we know the mean and standard deviation: $X = z\sigma + \mu$ And in a sample use: $X = zs + M$. Both formulas are functionally equivalent, because both formulas determine a raw score (X) by adding the mean of a distribution to the product of a z-score and the standard deviation of a distribution. From the example above, the minimum raw score needed for an A can be determined by insert the class' mean (*M* = 70), standard deviation (*s* = 10), and z-score (*z* = 2.5) into the formula above: X = (2.5)(10) + 70 = 95.

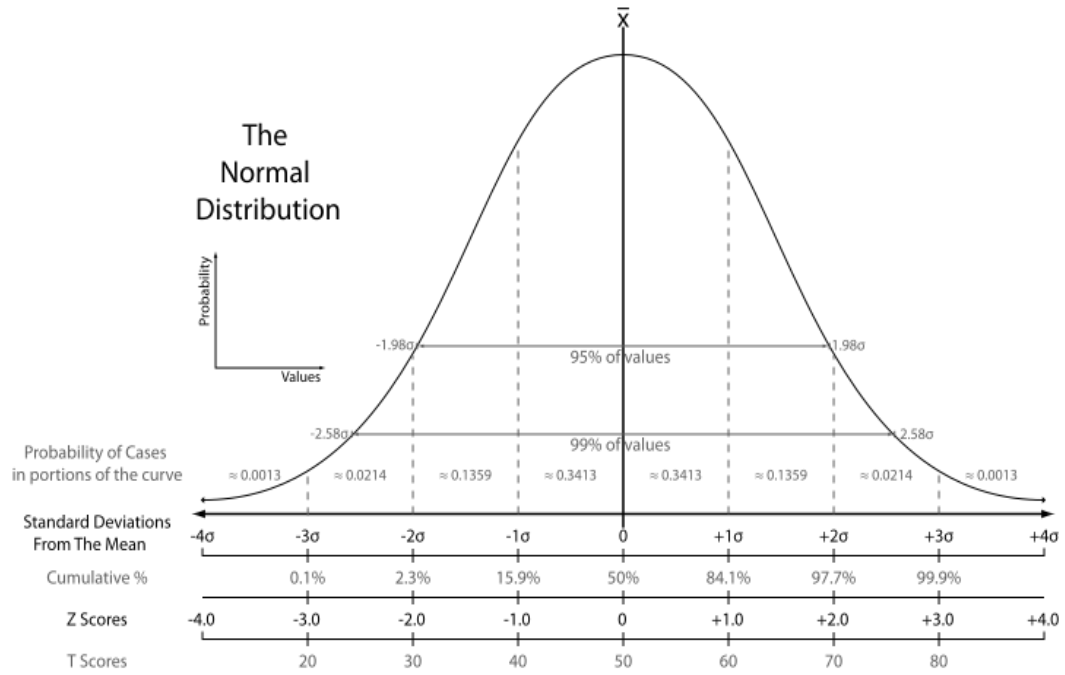
A score of X = 95 is exactly 2.5 standard deviations above the mean, so the minimum grade needed for an A is 95. What about a score that lies 1.25 standard deviations *below* the mean (z = -1.25)? Maybe this is

the minimum grade needed to pass the test. Determining this score is the same as before, you just need to be sure to take the negative z-score into account: X = (-1.25)(10) + 70 = 57.5. Thus, a score of 57.5 is exactly 1.25 standard deviations below the mean of 70, and is the minimum score needed to pass the test.

6.5 Standard Scores and the Normal Distribution

When you have a population or a sufficiently large sample of normally distributed data, a certain proportion of scores will fall between the mean and any given number of standard deviations from the mean. Specifically, the proportion of scores between the mean (μ or *M*) and one standard deviation from the mean is approximately .3413 (i.e., 34.13% of scores fall between the mean and one standard deviation from the mean, above or below). Similarly, the proportion of scores between one standard deviation and two standard deviations from the mean is approximately .1359. Incidentally, the proportion of scores between one standard deviation below the mean and one standard deviation above the mean is approximately .6826; hence, most scores in a normally distributed set of data fall right around the mean (i.e., central tendency).

These proportions can be seen in the general graph of the normal distribution, below. I should note this is a theoretical normal distribution centered on a mean and is symmetrical with half the scores greater than the mean and the half less than the mean. Remember, in a normal distribution most scores lie at the hump in the middle and fewer scores lie in the tails.



The height of the curve above any point in the x-axis reflects the probability of observing that particular score (X); hence, at any point along the x-axis, a certain proportion of scores are above and below that point. The higher the height of the curve, the more probable that score is likely to be observed and the greater the proportion of the distribution is made up of that score. As you can see in the graph above, as the distance between a score (X) and the mean increases, the height of the curve and the probability of observing that score decreases.

Along the x-axis, marks are given for the number of *Standard Deviations from The Mean*, from four standard deviations below the mean (-4σ) to four standard deviations above the mean (+4σ). Marks are also given to indicate the *Z Scores* relative to the mean from z = -4.0 to z = +4.0. (Don't worry about the *Cumulative %* and *T Scores*) As you can see, as you move farther and farther away from the mean in standard deviations, or z-Scores, the probability of being that far away become less and less. This fact about the normal distribution introduces two key points:

First, because the proportions/probabilities in the figure above are associated with specific numbers of standard deviations from the mean, and because standard deviations are a measure that can be used in any set of data, the proportions/probabilities associated with z-scores are universal and occur in any normally distributed data. For example, IQ scores are normally distributed with μ = 100 and σ = 16. This means the proportion of scores between an IQ score of X = 84 (μ - 1σ) and X = 116 (μ + 1σ) is approximately 0.6826. If we assume scores on the *Beck Depression Inventory* (BDI) are normally distributed with μ = 15 and σ = 5, the proportion of scores between a BDI score of X = 10 (μ - 1σ) and X = 20 (μ + 1σ) is also approximately 0.6826. Thus, regardless of the underlying data, as long as I is normally distributed these proportions/probabilities will be found for any z-score.

Second, by taking into account the mean and standard deviation of a distribution, we can calculate a z-score for any raw score and we can use the information about the normal distribution above to determine proportions and probabilities associated with that z-score, which we do in the next section.

6.6 Probabilities and Proportions in the Normal Distribution

Table 1 in Appendix A presents the is the **Standard Normal** (**z) Table**, which can be used to look up proportions of scores above or below z-Score within a normal distribution. A portion of Table 1 is presented below. There are three columns in Table 1, the first column lists z-Scores that increase from 0.00 to 4.00, mainly in increments of 0.01, and the second and third columns list proportions of scores under the normal curve that are between the z-score and the mean (column 2) and beyond that z-score and into the tail of the distribution (column 3).

Column 1	**Column 2**	**Column 3**	**Column 1**	**Column 2**	**Column 3**
z	**p (0 < x ≤ +z) or p (0 > x ≥ -z)**	**p (x ≥ +*z*) or p (x ≤ -*z*)**	***z***	**p (0 < x ≤ +z) or p (0 > x ≥ -z)**	**p (x ≥ +*z*) or p (x ≤ -*z*)**
1.00	0.3413	0.1587	**1.60**	0.4452	0.0548
1.01	0.3438	0.1562	**1.61**	0.4463	0.0537
1.02	0.3461	0.1539	**1.62**	0.4474	0.0526
1.03	0.3485	0.1515	**1.63**	0.4484	0.0516
1.04	0.3508	0.1492	**1.64**	0.4495	0.0505
1.05	0.3531	0.1469	**1.65**	0.4505	0.0495
1.06	0.3554	0.1446	**1.66**	0.4515	0.0485
1.07	0.3577	0.1423	**1.67**	0.4525	0.0475
1.08	0.3599	0.1401	**1.68**	0.4535	0.0465
1.09	0.3621	0.1379	**1.69**	0.4545	0.0455
1.10	0.3643	0.1357	**1.70**	0.4554	0.0446

Notice there are no negative z-Scores. This is because the sign (+/-) of a z-Score does not tell you anything the magnitude of a z-score r its distance from the mean; the sign only tells you the direction of the z-score relative to the mean. Thus, z = 1.25 is equal in distance to z = -1.25 and the proportions under the normal curve associated with a positive z-score will be the same as the proportions under the normal curve associated with a negative z-score of the same absolute value. If you have a negative z-score, just look up a positive z-score of the same absolute value and use those proportions in Table 1.

Before going into detail about what information columns 2 and 3 provide, let's set up an example. We have a normally distributed set of 19th Century American Literature test scores with μ = 80 and σ = 10. We calculate the z-score for the raw score of X = 92 and find it is z = 1.2. We now want to know the proportion of scores that lie between z = 1.2 and the mean of the distribution. Another way to conceptualize this question is we want to know the *probability* of obtaining a score between the mean and z = 1.2. Thus, we want to know p(μ ≤ X ≤ 92). The section of Table 1 that includes z = 1.2 is reproduced below with the area associated with z = 1.2 highlighted in yellow:

Column 1	**Column 2**	**Column 3**	**Column 1**	**Column 2**	**Column 3**
z	**p (0 < x ≤ +z) or p (0 > x ≥ -z)**	**p (x ≥ +*z*) or p (x ≤ -*z*)**	***z***	**p (0 < x ≤ +z) or p (0 > x ≥ -z)**	**p (x ≥ +*z*) or p (x ≤ -*z*)**

1.15	0.3749	0.1251	**1.75**	0.4599	0.0401
1.16	0.3770	0.1230	**1.76**	0.4608	0.0392
1.17	0.3790	0.1210	**1.77**	0.4616	0.0384
1.18	0.3810	0.1190	**1.78**	0.4625	0.0375
1.19	0.3830	0.1170	**1.79**	0.4633	0.0367
1.20	0.3849	0.1151	**1.80**	0.4641	0.0359
1.21	0.3869	0.1131	**1.81**	0.4649	0.0351
1.22	0.3888	0.1112	**1.82**	0.4656	0.0344
1.23	0.3907	0.1093	**1.83**	0.4664	0.0336
1.24	0.3925	0.1075	**1.84**	0.4671	0.0329

Column 2 in the Table provides the proportion of scores between the mean a given z-score; thus, the proportion of scores between the mean of μ = 80 and the raw score of X = 92 (z = 1.2) is 0.3849. This would be the same proportion of scores between the mean and z = -1.2, because the proportion of scores between the mean and a positive z-score is equal to the proportion of scores between the mean and a negative z-score of the same magnitude.

We can also use Table 1 to determine the proportion of scores beyond a given z-score, that is, the proportion of scores that fall in the tail of normally distributed. Another way to conceptualize this question is, what is the probability of having score greater than z = 1.2; thus, what is p(1.2 ≤ X). This proportion comes from Column 3 in the table above. For z = 1.2, the proportion of scores that fall in the tail 0.1151; hence, the probability of having a score greater than X = 92 (z = 1.2) is p = 0.1151. Note, the proportion of scores beyond z = 1.2 is equal to the proportion of scores in the tail for z = -1.2. Importantly, if the z-score is positive, this is the proportion of scores *greater* than the score (X), but if the z-score is negative, this proportion of scores *less* than the score.

The proportion of scores between the mean and any z-score (Column 2) to the proportion of scores that lie beyond that z-score (Column 2) add to 0.5000. This is because the proportion of scores above the mean in a normal distribution is always 0.5 and the proportion below the mean is always 0.5. This is so, because in a normal distribution the mean is equal to the median and there are always 50% of the scores above the median and 50% of the scores below the median.

Knowing this, you can use Table 1 to determine the probability of obtaining a score greater than or less than a given score (X). For example, with the same normally distributed set of 19th Century American Literature test scores (μ = 80; σ = 10), say we want to determine the probability of obtaining a score less than or equal to a raw score of X = 75. The z-Score for 75 in this distribution is z = -0.50. Look up z = 0.50 in Table 1, because there are no negative z-Values in the table. We are interested in the probability of obtaining a score less than or equal to X = 75 (z = -0.5), so we use the proportion in Column 3, which tells us that the probability of obtaining a score of 75 or less is p (X ≤ 75) = 0.3085. We use the proportion from Column 3 in Table 1, because that is the probability of being at that z-score or lower.

What about finding the probability of obtaining a score of X = 75 or more? In this case you use the same z-Score of z = -0.50 (again, use z = 0.50 in Table 1). There is no single proportion in Table 1 that provides you the answer, so what you do is find the probability of a score between the mean and z = 0.50, which is 0.1915 from Column 2, and add 0.5000. Adding 0.5000 takes into account the upper half of the distribution; that is, the probability of obtaining X = 75 or more is equal to the probability of obtaining a score between that raw score (X = 75) and the mean (μ = 80) added to the probability of obtaining a score in the other half of the distribution (above the mean). So, the probability of obtaining a score of 75 or greater is 0.6915.

How about the probability of obtaining a score of 90 or less? Again, calculate the z-score, which is z = 1.00. In this case, we need the probability below z = 1.00, which is equal to the area between the mean of the normal distribution and z = 1.00 (0.3413 from Column 2) added to the total area below the mean (0.5000). So, the probability of a score below 90 is 0.3413 + 0.5000 = 0.8413.

What if you wanted to calculate the probability of obtaining a score between two values? There are two ways to do this: (1) When both values fall on the same side of the mean, and (2) when one values falls above the mean and the other value falls below the mean.

From the same distribution of 19th Century American Literature test scores with μ = 80 and σ = 10, we want to know the probability of obtaining a score between 65 and 85. First, calculate the z-Scores for each raw score, which are z = -1.50 for 65 and z = 0.50 for 85. Next, look up the probability between each z-score and the mean in Column 2, which is 0.4332 for z = -1.50 and 0.1915 for z = 0.50. Because each score falls on a different side of the mean (one z-score was positive the other was negative), add these probabilities. So, the probability of obtaining a score between 65 and 85 is = 0.4332 + 0.1915 = 0.6247.

Say we want to determine the probability of obtaining a score between 85 and 90, which are on the same side of the mean. First, calculate the z-scores for each raw score, which are z = 0.50 for 85 and z = 1.00 for 90. Next, look up the probability between each z-Score and the mean in Column 2, which is 0.1915 for z = 0.50 and 0.3413 for z = 1.00. Because both scores are on the same side of the mean, subtract the smaller *probability* from the larger *probability*, which will leave you with the probability between 85 and 90. So, the probability of obtaining a score between 85 and 90 is 0.3413 - 0.1915 = 0.1498 (remember, probabilities cannot be negative).

The same procedure is used if both scores are less than the mean. For example, we want to know the probability of obtaining a score between 65 and 75, which are less than μ = 80. First, calculate the z-scores for each score (z = -1.50 for 65 and z = -0.50 for 75). Next, look up the probability between each z-score and the mean in Column 2 (0.4332 for z = -1.50 and 0.1915 for z = -0.50). Because both values are less than the mean, subtract the smaller probability from the larger probability. So, the probability of a score between 65 and 75 is 0.4332 - 0.1915 = 0.2417. Important: When calculating the probability between two scores that are on the same side of the mean, you will always subtract the smaller probability from the larger probability. But, when calculating the probability between two scores that are on different sides of the mean, add the two probabilities together.

CH 6 Homework Questions

1. Given a distribution with a mean of 15.565 and a standard deviation of 4.255, compute the standard score equivalents of the following scores:

a. X = 16.569 b. X = 10.895 c. X = 0 d. X = 15.565
e. X = 11.255 f. X = 20.525

2. Given a distribution with a mean of -5.000 and a standard deviation of 2.500, compute the standard score equivalents of the following scores:

a. X = -6.730 b. X = 2.950 c. X = -2.500 d. X = 0
e. X = -6.550 f. X = 0.850

3. Given a distribution with a mean of 10.000 and a standard deviation of 3.000, compute the raw score equivalents of the following standard scores:

a. z = 2.250 b. z = -1.000 c. z = -0.750 d. z = 0

4. If a person got a raw grade of 65 on a psychology test. Assuming that the raw grades will be curved based on the class' performance, which of the following class distributions would provide the most favorable interpretation of this raw grade? Why?

a. $\overline{X} = 55,\ s = 10$ b. $\overline{X} = 55,\ s = 5$ c. $\overline{X} = 60,\ s = 2.5$ d. $\overline{X} = 60,\ s = 1.5$

5. For the following distribution, convert a score of 40 to a standard score { 50, 45, 40, 40, 45, 50 }.

6. For the following distribution, convert a score of 50 to a standard score: 60, 55, 50, 50, 55, 60.

7. What are the values of the mean and the standard deviation for any set of standard scores?

8. What is a z-Score?

9. What proportion of z-scores in a normal distribution are:

a. -1.96 or more	b. -1.96 or less	c. 1.65 or less
d. 1.65 or more	e. 1.38 or less	f. -1.38 or more
g. between 0 and 2.00	h. between -2.00 and 0	i. between .40 and 3.01
j. between -3.01 and .40	k. between -1.24 and +1.24	

10. Suppose IQ scores in a population are normally distributed with $\mu = 100.00$ and $\sigma = 15.00$. What proportions of individuals have IQ scores of:

a. 100 or higher	b. 100 or less	c. between 110 and 120
d. 95 or less	e. 95 or higher	f. Between 90 and 110

11. Given a set of normally distributed scores with a mean of 100 and a standard deviation of 15, what score corresponds to a z-score of:

a. 2.75 b. 0 c. -2.00 d. -1.50 e. 1.25
f. 0.70

12. Given a normal distribution with a mean of 100 and a standard deviation of 15, what score would:

a. 33% of the cases be greater than or equal to?
b. 5% of the cases be greater than or equal to?
c. 2.5% of the cases be greater than or equal to?
d. 2.5% of the cases be less than or equal to?

13. Suppose that the income of a small town has a mean of $40,000.00 with a standard deviation of $10,000.00, what score would:

a. 33% of the cases be greater than or equal to?
b. 15% of the cases be greater than or equal to?
c. 2.5% of the cases be greater than or equal to?
d. 1.0% of the cases be greater than or equal to?
e. 0.1% of the cases be greater than or equal to?

14. Dr. Bob Ross is a humanist psychologist that wants to make people happy. He has developed a device called the Happy o'meter, which is used to measure the overall happiness of an individual. Scores on the Happy o'meter are normally distributed with $\mu = 25.250$ and $\sigma = 5.150$ Use these population parameters to calculate the standard (z) score for raw score (X) below:

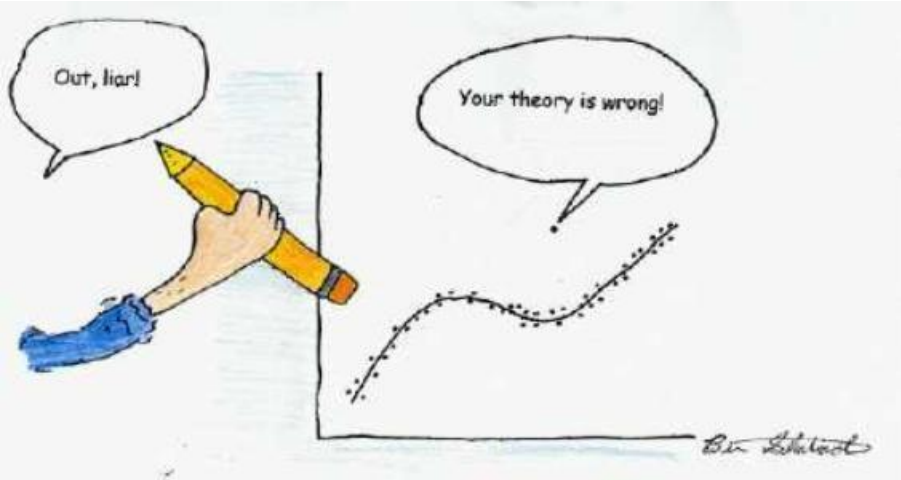
a. X = 28.350
b. X = 24.195
c. X = 31.565
d. X = 33.000
e. X = 20.250
f. X = 21.555
g. X = 35.255
h. X = 45.200

15 In a normally distributed population of scores with $\mu = 50$ and $\sigma = 15$, determine the proportion (not percentage) indicated in each question below. Please note that you will have to calculate the standard (z) score for each raw score first. Also, be sure to leave your proportions to four places (ten-thousandths place).

a. p (X > 55)
b. p (X < 75)
c. p (X < 25)
d. p (X > 35)
e. p (X > 45)
f. p (30 < X < 70)
g. p (55 < X < 65)
h. p (20 < X < 40)

16. A major form of identification in criminal investigations is fingerprints. Fingerprints vary on many dimensions, one of which is called the ridgecount. Suppose that you know that the ridgecounts of human beings follow a normal distribution with a mean of 165.00 and a standard deviation of 10.00. Suppose that a set of fingerprints was found at the scene of a crime and it was determined that the ridgecount was at least 200 (the exact value being in question because of smudging). Finally, suppose that a particular suspect has a ridgecount of 225. What should you conclude and why?

Chapter 7: Estimation and Sampling Distributions



7.1 Using Samples to Estimating the of Population

The picture above represents the concept of **estimation**, which is using sample data to estimate and make inferences about population parameters. Recall from Chapter 1, populations are theoretically infinitely large; hence, it is impossible to know the exact values of parameters with certainty, but we can use sample data to determine estimates of those parameters and infer what *would* likely be found in the population. So, we select a sample from a population and data from that sample to make inferences about what would be found in the population.

But, how good are sample statistics at estimating population parameters? How accurately does the sample mean (M), sample variance (s^2), and sample standard deviation (s) estimate the population mean (μ), population variance (σ^2), and population standard deviation (σ)? This chapter addresses how the sample mean is the unbiased estimator of the population mean, but sample variance and sample standard deviation are biased estimators, respectively, and how such bias is corrected.

7.2 Sample Mean is an Unbiased Estimator of the Population Mean

Assume you are a statistics professor and one of your a classes is populated with N = 25 students. On a quiz with range 0 – 10, you obtain the scores below. Below the scores is the true population mean (μ), variance (σ^2), and standard deviation (σ). If we consider this class to be a population, these are parameters, but assume you do not actually know these parameters.

10	10	9	9	8	8	7	7	6	6	6	6	5
5	5	5	4	4	4	3	3	2	1	1	1	

N = 25 $\quad$ $\mu = 5.4$ $\quad$ $\sigma^2 = 7.04$ $\quad$ $\sigma = 2.636$

You meet with other statistics teachers to compare grades. Remember, you do not know the parameters from your quiz, so you randomly sample $n = 5$ scores and calculate the sample mean, sample variance (s^2), and sample standard deviation (s). This sample contains the following scores: 10, 8, 5, 5, 3. The sample mean is $M = 6.2$, which is not equal to the population mean of $\mu = 5.4$. The difference of 0.8 between the sample mean and population mean is **sampling error** and is due to the sample not containing the entire population of scores. That is, any time you select a sample from a population, you are not selecting <u>all</u> the scores and all the individual differences in the population; you are selecting only part of that variability. Hence, sampling error will always present when working with sample data.

Say you randomly select another sample of *n* = 5 scores from the population of N = 25 quiz scores and this new sample contains scores 9, 6, 5, 4, 2. This second sample has a mean of *M* = 5.2, so the sampling error -0.2; thus, this sample's mean is closer to the population mean. If you were to continue this and generate more samples of *n* = 5 from this population, you would find some sample means are greater than μ, some are less than μ, and some are equal to μ, but the sample means will not be consistently greater than or less than the population mean. Thus the sample mean is an **unbiased estimator** of μ.

7.3 Sample Variance and Standard Deviation as Biased Estimators

If the sample mean is an unbiased estimator of μ, are the sample variance and sample standard deviation unbiased estimators of the population variance and population standard deviation? In the first sample of *n* = 5 from Section 7.2, the variance was $s^2 = 6.16$ and standard deviation was s = 2.482. In the second sample, the variance was $s^2 = 6.36$ and the standard deviation was s = 2.522. In both samples the variance and standard deviation were less than the population variance ($\sigma^2 = 7.04$) and standard deviation ($\sigma = 2.636$). If you continued randomly selecting samples from the *N* = 25 scores, you would find most sample variances and standard deviations are smaller than the population variance and standard deviation; thus, sample variance and sample standard deviation are **biased estimators**, because they underestimate the true population variance and population standard deviation.

When using a sample to estimate the population variance and population standard deviation, to correct for the underestimation divide the sum of squares (SS) in the sample by *n* – 1 rather than by *n*. This will produces an **estimate of the population variance**:

$$\hat{s}^2 = \frac{\Sigma(X-\overline{X})^2}{n-1} \quad \text{or} \quad \hat{s}^2 = \frac{SS}{n-1}$$

The hat (^) indicates the value is an *estimate* of the parameter. From Section 7.2, the first sample had a sum of squares equal to SS = 30.8, so the estimated variance is:

$$\hat{s}^2 = \frac{30.8}{5-1} = 7.7$$

The second sample in Section 7.2 had sum of squares equal to SS = 31.8, so estimated variance is:

$$\hat{s}^2 = \frac{31.8}{5-1} = 7.95$$

In both samples the variance estimates are slightly larger than the population variance ($\sigma^2 = 7.04$), but that's ok; you would rather overestimate variability than underestimate variability. Just note that if you used larger sample sizes to estimate the variance, the variance estimates would be better estimators.

From above, it follows that the **estimate of the population standard deviation** is equal to the positive square root of the estimate of the population variance:

$$\hat{s} = \sqrt{\frac{\Sigma(X-\overline{X})^2}{n-1}} \quad \text{or} \quad \hat{s} = \sqrt{\frac{SS}{n-1}} \quad \text{or} \quad \hat{s} = \sqrt{\hat{s}^2}$$

From samples 1 and 2 in Section 7.2 and above, we have:

Sample 1: $\hat{s} = \sqrt{7.7} = 2.775$ Sample 2: $\hat{s} = \sqrt{7.95} = 2.82$

Like the estimated variance, the estimated standard deviations from samples 1 and 2 are slightly larger than the population standard deviation (σ = 2.636). Again, this is ok, because you would rather overestimate variability than underestimate variability.

Importantly, if measuring the variability *within* a sample or *within* a population, use the sample formulas and the population formulas, respectively. That is, calculate sum of squares then divide by n (or N) to obtain the variance and take the square root of the variance to obtain the standard deviation. The only time SS is divided by n – 1 is when sample data is being used to *estimate* the population variance and standard deviation. This point is made more explicit in the table below, which lists the formulas you would use based on the type of variance measure that you are looking for:

Measure of Variability	Population	Sample	Population Estimate
Sum of Squares	$SS = \sum (X - \mu)^2$	$SS = \sum (X - \bar{X})^2$	$SS = \sum (X - \bar{X})^2$
Variance	$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$ --or-- $\sigma^2 = \frac{SS}{N}$	$s^2 = \frac{\sum (X - \bar{X})^2}{n}$ --or-- $s^2 = \frac{SS}{n}$	$\hat{\sigma}^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$ --or-- $\hat{\sigma}^2 = \frac{SS}{n - 1}$
Standard Deviation	$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$ --or-- $\sigma = \sqrt{\frac{SS}{N}}$ --or-- $\sigma = \sqrt{\sigma^2}$	$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n}}$ --or-- $s = \sqrt{\frac{SS}{n}}$ --or-- $s = \sqrt{s^2}$	$\hat{\sigma} = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}}$ --or-- $\hat{\sigma} = \sqrt{\frac{SS}{n - 1}}$ --or-- $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$

There are three important points regarding the table above. First, do not get bogged down in the symbols and nomenclature, focus on the concepts. Second, the measures of variability mean the same thing whether you are working within a population, sample, or using samples to estimate parameters. Specifically, the sum of squares measures total variability, variance measures average variability, and standard deviation is the average difference between a score and the mean. The only thing that differs between statistics, parameters, and parameter estimates is the group to which the measure refers. Finally, you should note the only difference in calculations across these measures is when calculating the variance estimate where you divide SS by n – 1 instead of by n. Very important: the sample variance and sample standard deviation are descriptive of the sample, whereas the population estimates that are based on the sample are descriptive estimated about what would be found in the population.

7.4 Degrees of Freedom

The n – 1 correction in the denominator is the **degrees of freedom** (***df***) in a sample, which is defined as the *number of scores that can vary independently in a sample*, that is, the number of scores in a sample that can take on any possible value, with one score being completely dependent on the other n – 1 scores. For example, say you have a sample with n = 5 scores and you know the mean is M = 10. Four (n – 1) of the five scores could be any value, but for the mean to be M = 10 the value of that last (fifth) score depends on the other n – 1 scores. Thus, four scores (n – 1) can vary independently from each other, but one completely score depends on the others. Still puzzled? Try this:

Say four (n – 1) of five scores are 8, 9, 11, 12. If M = 10, what must the fifth score be for the mean to be M = 10? The four scores sum to 40; hence, to obtain M = 10, the five (n) scores must add be 50. In this case, fifth score must be equal to 10. As another example, say four of the scores are 1, 2, 2, 5; what must the fifth score be for the mean to be M = 10?The sum of five scores must be equal to 50 to have M = 10, and because the sum of the four scores was only 10, the fifth score must be equal to 40 (50 – 10 = 40).

In any set of sample data, exactly n – 1 scores can vary independently of each other and theoretically take on any conceivable value for that variable. But exactly one score is completely dependent on the values of the other n – 1 scores. Thus, degrees of freedom are the number of scores allowed to vary independently with the last score being completely dependent on the others. In calculating estimates of the population variance and standard deviation the degrees of freedom determine the accuracy of those population estimates. That is, as degrees of freedom increase the accuracy estimating the population parameters also increases. This means that larger samples generate better estimates of the population.

7.5 Sampling Distribution of the Mean

The sample mean will be greater than or less than the population mean due to *sampling error* and the sample mean is always the best unbiased estimator of the population mean. Recall from Section 7.2, population of N = 25 quiz scores has a mean μ = 5.4, and the first sample's mean was M = 6.2 and the second sample's mean was M = 5.2; and both sample means were generated from a random n = 5 scores from the population. If we continued randomly selecting samples of n = 5 , with replacement, from the population of N = 25 scores so we eventually selected every possible sample of n = 5 scores from the population, we would have a collection of all possible samples from that population. (There are actually 25 x 25 x 25 x 25 x 25 = 25^5 = 9765625 possible samples of n = 5 scores.) If we calculated the mean of each sample, we would end up with a collection of all possible sample means of n = 5 from the population; hence, a **sampling distribution of the mean**.

A sampling distribution of the mean is a collection of all possible sample means of size n randomly selected with replacement from a population. For any population and for any given sample size, the sampling distribution of the means has the following characteristics:

1. The sampling distribution of the mean is normally distributed with the most frequently obtained (most probable) sample means being similar to the population mean. Hence, most sample means will be similar to the population mean.
2. The less frequent sample means are those that are discrepant from the population mean
3. The mean of the sampling distribution of the mean is centered on the population mean, that is, the mean of the sampling distribution of the mean (μ_M) is equal to the population mean (μ).
4. The standard deviation of the sampling distribution of the mean (σ_M) is equal to $\sigma/\sqrt{n}$, so long as all of the samples are the same size (n) and as long as $n < N$.

Points 3 and 4 form the basis of the **central limit theorem**, which states that *in any population with mean (μ) and standard deviation (σ); the sampling distribution of the mean for size n will have a mean equal to μ and standard deviation equal to* $\sigma/\sqrt{n}$*. The sampling distribution of the mean will approach a normal distribution as n approaches N.*

The standard deviation of the sampling distribution of means is more formally known as the **standard error of the mean** (**SEM**), which measures of the average distance between a population mean (μ) and a sample mean (*M*) of size n:

$$\sigma_{\overline{X}} = \sqrt{\frac{\sigma^2}{n}} \quad \text{or} \quad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

However, because parameters such as the standard deviation are often unknown, the standard error of the mean must often be *estimated* from sample data using the estimated standard deviation. The formula for the **estimated standard error of the mean** is:

$$\widehat{s}_{\overline{X}} = \frac{\widehat{s}}{\sqrt{n}} \quad \text{or} \quad \widehat{s}_{\overline{X}} = \sqrt{\frac{\widehat{s}^2}{n}}$$

The formula divides the estimated standard deviation by the square root of the sample size, and is our best estimate of the true standard error when the population standard deviation is unknown.

To demonstrate how the standard error of the mean is the standard deviation of the sampling distribution of the mean, and how the mean of the sampling distribution of the mean is equal to the population mean (μ), consider a hypothetical population of *N* = 4 {1, 2, 2, 3}. This population has mean μ = 2 and standard deviation σ = 0.707. I randomly select, with replacement, every possible sample of *n* = 2 scores from this population and calculate the mean of each sample. This is demonstrated in the table below. The X_1 and X_2 values are the scores in each possible randomly selected sample of *n* = 2. The table also shows the calculation of the sum of squares (each sample mean minus the mean of sampling distribution [i.e., 2]):

Sample	X_1	X_2	*M*	$(X - \mu_M)$	$(X - \mu_M)^2$
A	1	1	1.0	-1	1
B	1	2	1.5	-0.5	0.25
C	1	2	1.5	-0.5	0.25
D	1	3	2.0	0	0
E	2	1	1.5	-0.5	0.25
F	2	2	2.0	0	0
G	2	2	2.0	0	0
H	2	3	2.5	0.5	0.25
I	2	1	1.5	-0.5	0.25
J	2	2	2.0	0	0
K	2	2	2.0	0	0
L	2	3	2.5	0.5	0.25
M	3	1	2.0	0	0
N	3	2	2.5	0.5	0.25
O	3	2	2.5	0.5	0.25
P	3	3	3.0	1	1
N = 16			$\Sigma M = 32$ $\mu_M = 2$		SS = 4

The mean of the sampling distribution of the mean is $\mu_M = 2$, which is equal to the population mean for the N = 4 scores (μ = 2). The standard deviation of these N = 16 sample means is:

$$\sigma_M = \sqrt{\frac{SS}{N}} = \sqrt{\frac{4}{16}} = \sqrt{0.25} = 0.5$$

If we calculate the standard error of the mean using the population standard deviation (σ = 0.707) and sample sizes (n = 2), we get:

$$\sigma_{\overline{X}} = \frac{0.707}{\sqrt{2}} = 0.5$$

This is equal to the standard deviation of the sampling distribution of the mean (0.5); hnce, the standard error of the mean is the standard deviation of the sampling distribution of the mean. Remember, a standard deviation measures the average difference between a score (X) and the mean. Thus, the standard error of

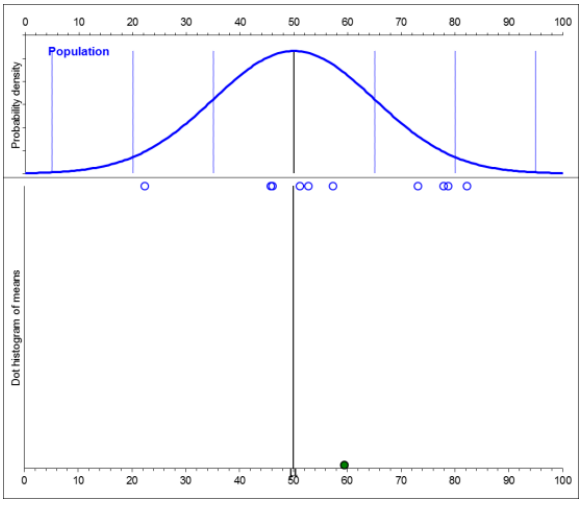
the mean measures the average difference between a sample mean (*M*) of size *n* and the population mean (μ). In some sense than, the standard error of the mean measures the average, or expected, sampling error for a given sample size.

The standard error of the mean is influenced by the population standard deviation (*σ*) and the sample size (n). Holding sample size constant, greater population variance (larger σ) results in a larger a standard error of the mean. Holding population variance (σ) constant larger sample sizes result in smaller standard error. That the standard error of the mean is influenced by the sample size demonstrates an important characteristic of the central limit theorem: as n approaches N the sampling distribution of the mean approaches a normal distribution. Another way to state this is that as n gets larger the sampling distribution of the mean becomes more "normal." Or, more succinctly, larger sample sizes produce better approximations of the population. This means that by using a larger sample size, the population estimates that come from a sample will more accurately estimate the parameters. The standard error of the mean will decrease as sample size increases. Thus, as sample size increases and the standard error decreases and sample means will be on average closer estimators of μ, which reflects more accuracy in the sample estimating the population.

7.6 The "Dance of the Mean" and the "Mean Heap"

The title of this section is inspired by topics of the same name in Geoff Cumming's (2012) recent book on issues in modern statistics. This point of this section is to follow up on some issues in the preceding sections regarding sampling distributions and what happens when you sample from large populations. In particular, how sample means are unknown in infinitely large populations. When randomly selecting a sample of size *n* from a population, even if the population mean and standard deviation are known, which is rare, you never can be certain where the sample mean will fall relative to the population mean. Thus the value of the population mean can be thought of as a random variable that has an unknown value prior to data collection.

Consider a population with μ = 50 and σ = 15, but an unknown size (i.e., N is infinitely large and unknown). I randomly select a sample of *n* = 10 scores, which are represented as the blue circles along the x-axis below the population distribution in the figure to the right.[1] This sample of *n* = 10 has a mean *M* = 58.76, which is represented by the green circle at the bottom; thus, the sampling error was 58.76 – 50 = 8.76.

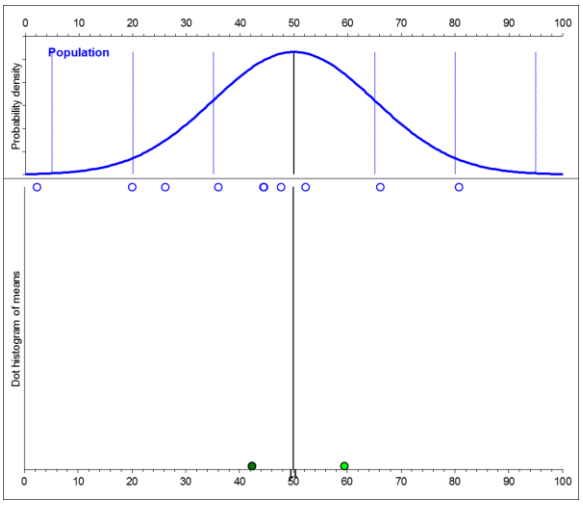


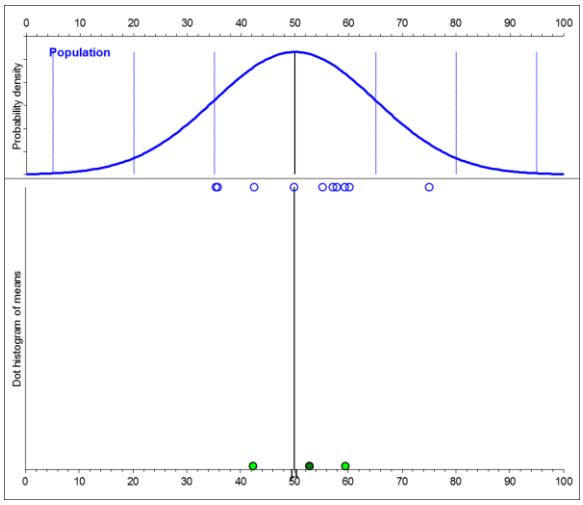
If I took a second sample (left figure, below) and a third sample (right figure, below) from the same population, we might get the following samples and means, below. In each figure, the data in each new sample (blue dots) appears under the population curve and the newest sample's mean appears at the bottom (in green); and the mean for the second sample was *M* = 42.24 and the mean for the third sample was *M* = 52.80.

These three samples demonstrate the variability in the sample means or, what Cumming (2012) calls, the **dance of the means**. Simply, the dance of the means shows that sample means based on randomly selected samples from a population will "dance" around the population mean; thus, there is variability across sample means taken from the same population. Perhaps more importantly, this means you can never be

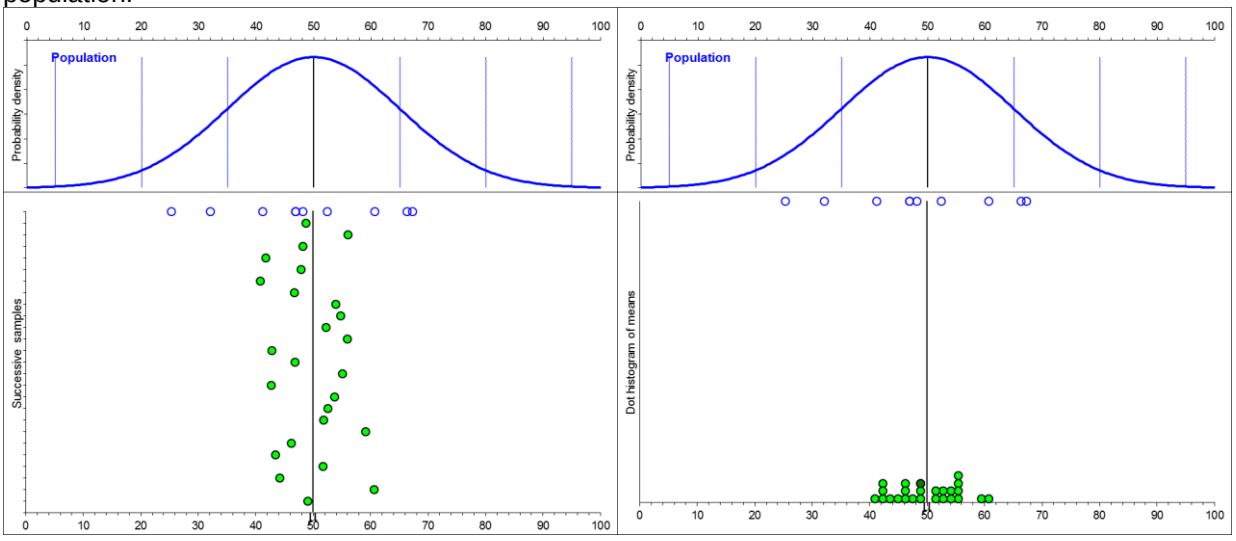
[1] The figures in this section were created using ESCI, which can be obtained from Geoff Cumming's website: http://www.latrobe.edu.au/psy/research/cognitive-and-developmental-psychology/esci

100% sure about the value of a sample mean until the data are collected, because the data for the sample are randomly selected; hence, now two sample means are alike (like snowflakes!).

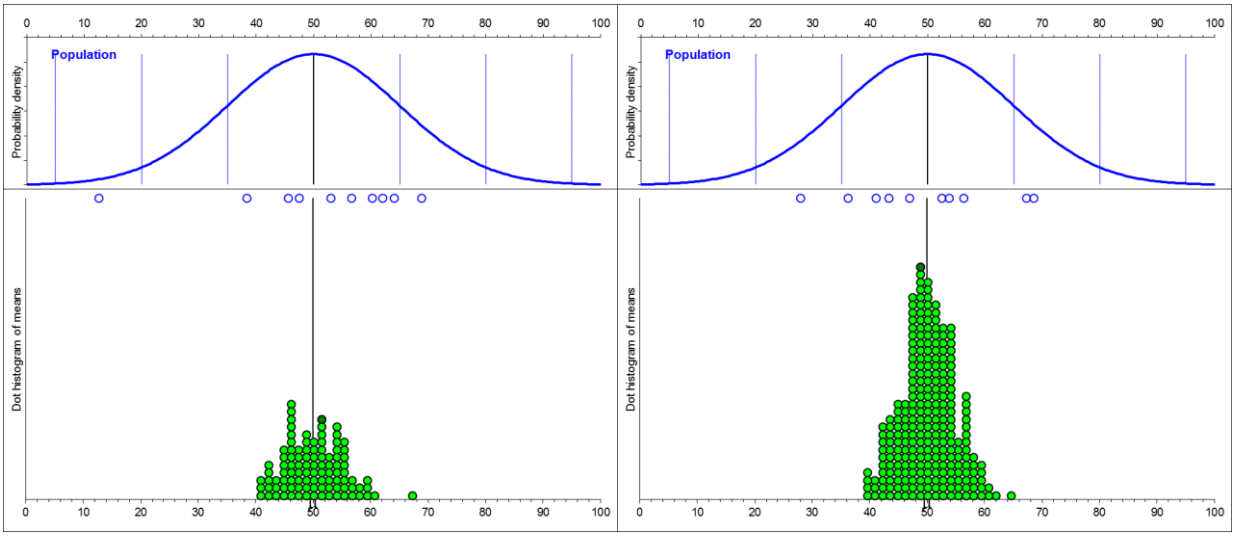




The figure below, on the left, illustrates this dance of the means more, with 25 sample means "dancing" around μ = 50 (the blue dots are for the most recently-selected sample). The figure below, on the right, illustrates another concept called the **mean heap**, which is simply a pile of the sample means. What you should see in the mean heap is that means with the same value are piled on top of each other; hence, the mean heap creates a frequency histogram for sample means based on randomly-selected samples from a population.



Below, two additional mean heaps are presented from the same population. The figure on the left is a mean heap for 100 sample means, and the figure on the right is a mean heap for 250 sample means. Notice that although the sample means dance around the population mean, they tend to form a mean heap centered on the population mean, that is, the most frequency sample means are equivalent to or nearly equivalent to the population mean, and the less frequent means are more discrepant from the population mean. Also notice that the man heap appears to take on the shape of a normal distribution as the number of samples increases.



Thus, the dance of the mean and the mean heap illustrate several important concepts covered earlier. First, the mean is an unbiased estimator of the population mean. This comes from the dance of the means and the fact that some sample means are greater than μ, some are less than μ, and some are equal to μ. Second, the mean is the best estimate we have of the population mean. This is because of point one and the fact that the most frequent sample means "pile" around μ. Lastly, because the distribution of sample means approximates a normal distribution, we can use characteristics of the normal distribution to make inferences about a normally-distributed population from our samples. However, one question is, just how good does the sample mean estimate the population mean (μ)? To answer that, we turn to confidence intervals.

7.7 Confidence Intervals around a Sample Mean

So now you've collected data, calculated the sample mean, which is your best estimate of μ, and you have estimated population variance and standard deviation using your sample data. Your sample mean may be different from the population mean due to sampling error; indeed, the dance of the means in Section 7.6 illustrated how sample means randomly vary around the population mean. Additionally, because population means are not always known, it would be nice to know approximately where the population mean (and other sample means) might fall relative to a single sample mean. To do this, we establish **confidence intervals** around the sample mean.

The confidence interval is a range around a sample mean (not around the population mean) of plausible values for μ, thus, it is how precise our estimation of μ is based on our sample. Stated differently, a confidence interval is a *range around a sample mean that has a certain likelihood of including an unknown population parameter*. The question is, how probable/likely do we want to be? Researchers generally calculate a **95% confidence interval** (**95% CI**) or ***99% confidence interval*** (**99% CI**) around the sample mean, and this percentage can be interpreted in multiple ways.

However, before getting into confidence intervals, I should mention the **margin of error** (**MOE**), which is the largest likely (probable) estimation error (sampling error) between μ and *M*. Hence, MOE is the largest likely difference we should expect between a sample mean (*M*) and a population mean (μ), whether that population mean is known or unknown. Stated differently, the MOE is a range of values that likely contains the population mean (as you will see below, MOE is one-half the confidence interval). Because we want to be very certain a sample mean is a good estimator of the population mean, we define the likelihood that the MOE contains the population to be 95% or greater. That is, we would be 95% certain that the MOE is the largest estimation error between μ and *M*, and we would be 95% certain μ is within the MOE. (You can

calculate MOEs or any likelihood, but 95% is the smallest one normally computes). The formula for the MOE is:

$$MOE = z_{\alpha}\sigma_{\overline{X}}$$

In the formula, $\sigma_{\overline{X}}$ is the standard error of the mean and z_{α} comes from the standard normal table (Table 1) in Appendix A and is based on the likelihood you choose for the MOE (e.g., 95%). As an example of the MOE, let's return to the statistics quiz examples form Section 7.2, where we had a population of quiz scores with μ = 5.4 and σ = 2.636. Two samples each had *n* = 5 scores, and the mean for sample 1 was *M* = 6.2 and the mean for sample 2 was *M* = 5.2. For each sample, because *n* is the same and comes from the same population, the standard error of the mean will be the same:

$$\sigma_{\overline{X}} = \frac{2.636}{\sqrt{5}} = \frac{2.636}{2.236} = 1.179$$

The value for z_{α} is associated with the likelihood chosen for the MOE (i.e., 95%) and is a z-score from Table 1 in Appendix A. Note that the MOE should include all but the most extreme or distant 5% of the scores (100% - 95% = 5%) in a distribution around a sample mean. This 5%, when expressed as a proportion (.05), is the **alpha level** (**α**). The alpha level is the probability a population mean will not fall within the MOE (we'll discuss the alpha-level in more detail in later chapters).

The value for z_{α} is found by looking up half the alpha-level in Column 3 of Table 1. That is, divide your alpha-level in half (.05/2 = .025) and look that value up in column 3 (you divide the alpha-level in half, because columns 2 and 3 in Table 1 deal with half of the distribution around a mean). Ponce you locate that value in column 3, use the z-score in column 1 for z_{α}. When you look up α/2 = .025 in column 3 of Table 1, you should find a z-Score equal to 1.96. This is z_{α} for the MOE that we can use for both of the samples:

$$MOE = 1.96 \times 1.179 = 2.311$$

This tells us the largest likely estimation error (difference) between μ and *M* is 2.311. Stated differently, we are 95% certain the population mean will not differ from the sample mean (*M*) by more than 2.311 points. That is, we are 95% certain the largest difference between μ and *M* will be 2.311 points.

So what about the confidence interval? The **confidence interval** (**CI**) is not much different than the MOE and, indeed, CIs actually can use the MOE. Earlier, I indicated that the MOE is one half of the CI around a sample mean. Using the example above, the largest estimation error we should expect between μ and *M* is +2.311 (sample mean is less than μ) or -2.311 (sample mean is greater than μ). Thus, the CI designates a range around (on both sizes) of a sample mean:

$$CI_{\alpha} = \overline{X} \pm z_{\alpha}\sigma_{\overline{X}} \qquad \text{or} \qquad CI_{\alpha} = \overline{X} \pm MOE$$

The CI is simply the MOE added to and subtracted from around a sample. We can use this formula and the information from the two samples above to determine the CIs around the sample means:

Sample 1: Lower Limit [LL] = 6.2 - 2.311 = 3.889 and Upper Limit [UL] = 6.2 + 2.311 = 8.511
Sample 2: Lower Limit [LL] = 5.2 - 2.311 = 2.889 and Upper Limit [UL] = 5.2 + 2.311 = 7.511

These values are the boundaries of the 95% CI around each sample mean. Generally, the CI around a sample mean is reported as [LL,UL]; so for sample 1 the CI would be reported as [3.889, 8.511] and for sample 2 would be reported as [2.889, 7.511]. But what does the CI mean? Table 5.1 in Cumming (2012) provides six different uses of the CI, but the following two are important for present purposes:

1. The CI is one of many possible CIS that can result from randomly selecting samples from a population. That is, the dance of the means shows us that values of the sample mean randomly vary around the population mean, if it is known; hence, the limits of the CI will also randomly vary.

Given this, over the long run 95% of CIs will contain the population mean and 5% will not contain the population mean.

2. The Ci in the form [LL,UL] represents the range of most plausible values for μ. How plausible/likely does the CI contain μ? If we construct the 95% CI around *m*, then we can be 95% certain μfalls within this range of plausible values.

7.8 Confidence Intervals with Unknown Parameters

If the population standard deviation and mean are unknown, you cannot calculate the standard error of the mean (σ_M), but you can estimate the standard error of the mean and use it to estimate the confidence interval around the sample mean. Recall, the estimated standard error of the mean is:

$$\widehat{s}_{\overline{X}} = \frac{\widehat{s}}{\sqrt{n}}$$

What about z_α? Because we do not know the population parameters, using z-scores is inappropriate because we do not know anything about the population distribution, so we to estimate z_α. This estimate of z_α, which we call t_α, comes from the ***t*-Distribution**, which is presented in Table 2 in Appendix A. A portion of the *t*-tables is below, which is described in more detail in later chapters.

The t-tables are similar to the z-tables: The leftmost column lists *t*-Values and the values in the table are probabilities under the t-Distribution. The column headings list degrees of freedom (df) values; hence, to find t_α you need the degrees of freedom in the sample. Each sample from Section 7.2 included $n = 5$ scores, hence, the degrees of freedom in each sample was df = 4, which is highlighted in yellow below.

Table 2: Probabilities Under the *t*-Distribution

	df														
t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2.71	.1125	.0567	.0366	.0268	.0211	.0176	.0151	.0133	.0120	.0110	.0101	.0095	.0089	.0085	.0081
2.72	.1121	.0564	.0363	.0265	.0209	.0173	.0149	.0131	.0118	.0108	.0100	.0093	.0088	.0083	.0079
2.73	.1118	.0560	.0360	.0262	.0206	.0171	.0147	.0129	.0116	.0106	.0098	.0091	.0086	.0081	.0077
2.74	.1114	.0557	.0357	.0260	.0204	.0169	.0145	.0127	.0114	.0104	.0096	.0090	.0084	.0080	.0076
2.75	.1110	.0554	.0354	.0257	.0202	.0166	.0143	.0125	.0112	.0102	.0094	.0088	.0083	.0078	.0074
2.76	.1106	.0550	.0351	.0254	.0199	.0164	.0140	.0123	.0111	.0101	.0093	.0086	.0081	.0077	.0073
2.77	.1103	.0547	.0348	.0252	.0197	.0162	.0138	.0121	.0109	.0099	.0091	.0085	.0080	.0075	.0071
2.78	.1099	.0543	.0345	.0249	.0195	.0160	.0136	.0120	.0107	.0097	.0090	.0083	.0078	.0074	.0070
2.79	.1095	.0540	.0342	.0247	.0192	.0158	.0135	.0118	.0105	.0096	.0088	.0082	.0077	.0072	.0069
2.80	.1092	.0537	.0339	.0244	.0190	.0156	.0133	.0116	.0104	.0094	.0086	.0080	.0075	.0071	.0067

Once you locate the correct column, scroll down that column until you find the probability associated with half of the alpha level (i.e., .05/2 = .025). In this case, because .025 does not appear in the table exactly, locate the next smallest probability (.0249), and then the value in the leftmost column (2.78) will be t_α. The expression for estimating the confidence interval with unknown population parameters is:

$$CI_\alpha = \overline{X} \pm t_\alpha \widehat{s}_{\overline{X}}$$

The is identical to the formula in Section 7.6, except that in place of the standard error of the mean, the estimated standard error of the mean is being used, and in place of z_α, t_α is used. The estimated standard deviations for samples 1 and 2 in Section 7.2 are 2.775 and 2.820, respectively. The estimated standard

error of the mean for sample 1 is 1.241, and for sample 2 is 1.261. Thus, the 95% MOE for each sample is:

$$\text{Sample 1: } \square\square\square = \pm\square_{\square}\hat{\square}_{\underline{\square}} = 2.78 * 1.241 = 3.467$$
$$\text{Sample 2: } \square\square\square = \pm\square_{\square}\hat{\square}_{\underline{\square}} = 2.78 * 1.261 = 3.506$$

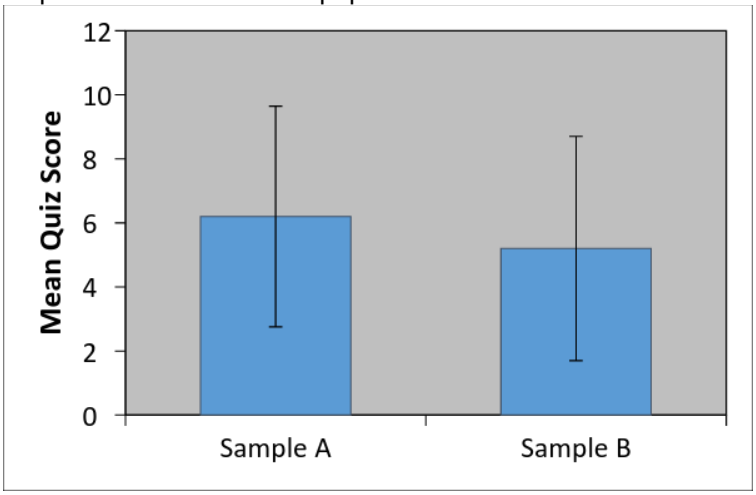
And the confidence intervals and limits are:

$$\text{Sample 1: } \square\square_{\square} = \underline{\square} \pm \square_{\square}\hat{\square}_{\underline{\square}} = 6.2 \pm 3.467 \qquad [2.773, 9.667]$$
$$\text{Sample 2: } \square\square_{\square} = \underline{\square} \pm \square_{\square}\hat{\square}_{\underline{\square}} = 5.2 \pm 3.506 \qquad [1.694, 8.706]$$

Like the confidence intervals calculated in Section 7.7, these confidence intervals are ranges of plausible values from the population mean. Stated differently, we can be 95% certain the population mean is contained within the confidence interval.

7.9 Displaying Confidence Intervals in Graphs

Confidence intervals are often included in graphs that display means, by having the graphing program draw a line from mean to the UL and from the mean to LL. For example, the graph to the right depicts the means from samples 1 and 2 from the earlier sections. The height of each bar represents the mean of each sample and the black line extending above the mean and below the mean represents the endpoints (confidence limits) of the 95% confidence interval around the sample means. This is a common way to graphically present where the most plausible values for the population mean.



There are very simple ways to do this in Excel and other graphing programs. Importantly, APA formatting suggests including confidence intervals in all graphs that present numerical data, so it is good to become familiar with how to create them.

7.10 Using CIs to Determine Difference from a Null Value

Recall from Chapter 1, the null hypothesis (H_0) which predicts no difference or no relationship exists between variables, and the alternate hypothesis (H_1), which predicts there is a difference or a relationship between variables.

Chapter 10 introduces the concept of null hypothesis testing in detail, but here I would like to discuss how confidence intervals can be used to determine whether a sample mean is statistically different from a value

predicted by the null hypothesis, by examining whether a value predicted by the null hypothesis is within the confidence interval. This is an important concept in statistics, because most inferential statistics are used to test the predictions of null hypotheses, often by examining whether two means are different from one another.

To introduce this idea of using confidence intervals to evaluate null hypotheses, let's set up an example. Assume you are a gardener and every year you grow green beans in your backyard garden, and every year you measure the length of each green bean you pick. Because you measure every bean and the beans come only from your backyard garden, you could consider this a population of data; hence, you know that the mean beam length is μ = 8.0 cm with σ = 2.0 cm.

One year, a botanist friend of yours tells you she has developed "magical bean growing juice" that will increase the length of beans and also make them crispier and tastier. She is going to sell the magical bean juice in the future, but she gives you a sample so you can use it on your bean crop this year. You really like beans, so you decide to try out the magical bean juice on some of your plants; if the bean juice works and makes your beans longer, you'll buy some in the future. If it doesn't work, you won't buy the magical bean juice. The question is, how will you know the magical bean juice will work?

You decide to use the bean juice on just a few of your bean plants; hence, you'll have a sample of beans that are given the magical bean juice. If the bean does not work, you would assume that the average length of the beans in this sample will not be much different than the average of all of your beans. Recall, the null hypothesis generally predicts no difference will be found, so in this case the null would predict the bean juice will not increase the length of the beans or, stated differently, the null hypothesis would predict the length of the beans in this sample should be the same as the length of the beans in your population of beans. Symbolically, this can be written as:

H_0: μ = 8.0 cm

That is, the null predicts that you will find a sample mean that is equal to the known population mean. In contrast, the alternate hypothesis generally predicts that a difference will be found, so in this case the alternate would predict the bean juice *will* increase the length of the beans or, stated differently, the alternate hypothesis predicts the length of the beans in this sample should be different from the length of the beans in your population of beans. Symbolically, this can be written as:

H_1: μ ≠ 8.0 cm

That is, the alternate predicts you will find a sample mean that is unequal to the known population mean. You run your bean juice study and end up picking 100 beans that were given the magical bean juice. Your sample has a mean of *M* = 8.5 cm, so it looks like the bean juice worked. However, it is important to remember the "dance of the means," that is, when randomly selecting subjects (beans) and measuring those subjects (bean lengths), the sample mean will unpredictably be above, below, to equal to the population mean. Hence, it is possible the bean juice did <u>not</u> actually work and you just happened to get a mean of *M* = 8.5 cm, by random selection. Thus, you need to determine how <u>plausible</u> this sample mean is if the bean juice did not actually work and this is where the confidence intervals come into play.

You decide to calculate the 95% CI around the sample mean of *M* = 8.5 cm. The standard error of the mean based on your sample o *n* = 100 beans is:

$$\underline{\sigma}_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2.0}{\sqrt{100}} = \frac{2.0}{10} = 0.20$$

Using z_α = 1.96, the 95% CI around the sample mean is:

$$CI_{95} = \underline{\bar{X}} \pm z_\alpha \underline{\sigma}_{\bar{X}} = 8.5 \pm 1.96 * 0.20 = 8.5 \pm 0.392$$

And the lower and upper limits of the confidence interval around the mean are: LL = 8.5 – 0.392 = 8.108 and UL = 8.5 + 0.392 = 8.892; hence, the sample mean and CI is 8.5 [8.108, 8.892].

Remember, the CI is the range of plausible values for a population mean; hence, it is the areas around a sample mean we are 95% certain the population mean falls. Also remember, the null hypothesis predicts the sample mean should be equal to the overall population mean of bean length (H_0: μ = 8.0 cm). But in this example, the population mean is not contained within the CI around the sample mean and actually falls outside of the CI by 8.108 – 8.0 = 0.108 cm.

When the mean that is specified or predicted by the null hypothesis does not fall within the CI around the sample mean, the null hypothesis is "***rejected***". Rejecting the null hypothesis means because the value it predicts is not one of the plausible values around the sample mean, we should not accept the prediction of the null hypothesis. And in this case, we reject the null, and we "***accept***" the alternate hypothesis, because the alternate hypothesis predicts the sample you obtain will be different from the population mean, that is, the CI around the sample mean will not include the population mean. Because we reject the null hypothesis, we would conclude the magical bean juice works!

But what if the sample mean had been *M* = 8.20? In this case (assuming you had picked *n* = 100 beans, the CI around the sample mean would be:

$$\text{CI}_{95} = \bar{X} \pm t_{95} s_{\bar{X}} = 8.2 \pm 1.96 * 0.20 = 8.2 \pm 0.392$$

And the lower and upper limits of the confidence interval around the mean are: LL = 8.2 – 0.392 = 7.808 and UL = 8.2 + 0.392 = 8.592 hence, the sample mean and CI is 8.2 [7.808, 8.592]. In this case, the CI contains the population mean specified by the null hypothesis (H_0: μ = 8.0 cm); hence, the population mean is one of the plausible values in the 95% CI. When this happens, you "***retain***" the null hypothesis, because the range of values you obtained includes plausible value if the bean juice did not actually work.

CH 7 Homework Questions

1. What is sampling error? What is sampling error due to? How can we represent the amount of sampling error that is present in a statistic?

2. What is an unbiased estimator? What is a biased estimator? Which sample statistics are unbiased and which are biased?

3. Why is the sample mean the "best estimate" of a population mean?

4. Why is the sum of squares divided by n - 1 rather than by n when computing the variance estimate?

5. What are degrees of freedom?

6. What is the sampling distribution of the mean?

7. What characteristics of a sampling distribution of the mean are addressed by the central limit theorem?

8. What will the mean of a sampling distribution of the mean always equal? Why?

9. What is the standard error of the mean? What information does it convey? How is it calculated?

10. What is the difference between the standard error of the mean and a standard deviation of a set of raw scores?

11. In cases where the population standard deviation is unknown, or does not exist, the standard of the mean cannot be calculated. What is used in its place, and how is it calculated?

12. What are the two factors that influence the size of the standard error of the mean, and how do they do so?

13. Say that a sampling distribution of the mean has a standard error of the mean that is equal to 0. What does this indicate about the means of the samples drawn from the population?

14. Say that a sampling distribution of the mean has a standard error of the mean that is equal to 0. What does this say about the variability of scores in the population (σ)?

15. Compute the variance and the standard deviation for the following scores: { 5, 4, 2, 6, 3, 7, 3, 1, 4, 5 }. Next, estimate the variance and estimate the standard deviation in the population, and then compare the two sets of results.

16. Compute the variance and the standard deviation for the following scores: 6, 5, 3, 7, 4, 8, 4, 2, 5, 6, 3, 6, 5, 4, 7, 5. Estimate the variance and the standard deviation in the population, and compare the two sets of results.

17. Compute the variance and the standard deviation for the following scores: { 7, 6, 8, 9, 7, 6, 7, 8, 4, 8 }. Next, estimate the variance and estimate the standard deviation in the population, and then compare the two sets of results.

18. The following scores below represent a sample of working memory capacity scores taken from a group of research subjects in a study on memory

62	73	73	53	74	55	66	49
64	74	75	65	68	75	64	

a. Calculate the sample mean.
b. Calculate the sum of squares.
c. Calculate the sample variance.
d. Calculate the sample standard deviation.
e. Calculate the estimated population variance.
f. Calculate the estimated population standard deviation.

19. Say that the population standard deviation for working memory capacity scores is σ = 8. Calculate the standard error of the mean based on the sample size from #18.

20. Calculate the 95% margin of error and the upper and lower boundaries for the 95% confidence interval around the sample mean in exercise #18, using the standard error of the mean from #19.

21. Calculate the 99% margin of error and the upper and lower boundaries for the 99% confidence interval around the sample mean in exercise #18, using the standard error of the mean from #19.

22. *Use the following to answer the questions below*. The "attentional control scale" is a device used for measuring individual differences in a person's ability to control and maintain their focus of attention and thinking. Scores on the attentional control scale range from 15 to 60, with higher scores reflecting better

control over attention. Assume that scores on the attentional control scale are normally distributed with μ = 37.5 and σ = 10. You administer the attentional control scale to a random sample of twelve psychology majors and obtain the scores listed below. Use them to answer the following questions:

26 28 47 21 38 27 38 18 36 22 26 45

a. Calculate the sample mean.
b. Calculate the sum of squares.
c. Calculate the estimated population variance.
d. Calculate the estimated population standard deviation.
e. Calculate the standard error of the mean, based on the population standard deviation.
f. Calculate the 95% confidence interval around the sample mean.
g. Calculate the estimated standard error of the mean, based on the estimated standard deviation.
h. Using the estimated standard error of the mean, calculate the 95% confidence interval around the sample mean.

23. *Use the following to answer the questions below*. Short term memory is the storage area of memory that maintains information in an active state just long enough for that information to be moved and stored in long term memory. The capacity of short term memory is generally μ = 7 pieces of information and σ = 2. A short researcher tests the short term memory for a sample of five neuroscience majors and obtains the following short term memory scores. Use this information to answer the questions below:

6 8 9 8 9

a. Calculate the sample mean.
b. Calculate the sum of squares.
c. Calculate the estimated population variance.
d. Calculate the estimated population standard deviation.
e. Calculate the standard error of the mean, based on the population standard deviation.
f. Calculate the 95% confidence interval around the sample mean.
g. Calculate the estimated standard error of the mean, based on the estimated standard deviation.
h. Using the estimated standard error of the mean, calculate the 95% confidence interval around the sample mean.

24. What two things does the size of the confidence interval reflect?

25. A sample of n = 20 is drawn from each of two populations. For population A, μ = 8.00 and σ = 4.00; and for population B, μ = 9.00 and σ = 6.00. Which sample mean is likely a better estimate of its population mean? Why?

26. Compute the estimated standard error of the mean for the data in Exercise 15.

27. Compute the estimated standard error of the mean for the data in Exercise 16.

28. Compare the estimated standard error of the mean calculated in Exercises 26 and 27. Which sample mean is probably a better estimate of its population mean? Why?

29. A population has the following parameters: μ = 500 and σ = 100. A sample of n = 100 subjects is randomly selected, and has *M* = 520. Calculate the 95% confidence interval around the sample mean.

30. Calculate the 99% confidence interval based on the information in #29.

31. Say that the sample size from #29 and 30 was increased to n = 1000. Recalculate the 99% confidence interval based on the information in #29 and 30.

Chapter 8: Probability

8.1 The Monty Hall Paradox

I've never understood the reason, but humans hate probabilities. *Math is power*...*math is power*! I think a lot of this comes from not taking the time necessary to evaluate information and just calculate the first probability that comes to mind. You should always evaluate the situation and conceptualize what probability needs to be calculated before putting pencil to your paper. The following is a good example of people failing to take information into account when estimating probabilities.

On the game show "Let's Make a Deal" with its host, Monty Hall, one contestant round would go to the prize round, where the contestant is presented with three doors, Door 1, Door 2, Door 3. Behind one door was a fabulous prize, such as a new car, money, or a vacation, and behind the other two doors were gag prizes such as goats. Monty Hall tells the contestant to pick one door. At this point, the probability of correctly selecting the door with the prize 1/3, because there are three doors and only one door has the prize. Assume Door 1 was initially selected by the contestant. Monty Hall then says, '*I'll tell you what I'm going do*. *I'll open one of the two remaining doors, you can then decide to stay with your original choice or switch to the other door*'. Assume Monty Hall opens Door 2, revealing a goat...and this is where the fun begins...

The **Monty Hall Paradox** boils down to whether you should stay with your original choice (Door 1) or switch to the unopened door (Door 3). Are you more likely to win by staying with your original choice of Door 1 or are you more likely to win by switching to the other unopened door (Door 3)? When presented with this problem most people assume it doesn't matter whether you stay or switch, because most think the probability of winning by staying is 0.5 and the probability of winning by switching is 0.5. That is, most assume the probability of winning by switching and winning by staying are both 50/50, because there are now only two doors and the prize is behind one door.

And here's the catch: the actual probability of winning by staying is 1/3 and the probability of winning by switching is 2/3; hence, you have a better chance of winning if you switch from your original choice! We'll go over the proof of this later, but for now think about this: the reason people assume the probability of winning by staying and by switching are equal is because people neglect the fact that Monty Hall knows where the prize is! When Monty Hall opens a door he provides information about the prize location. This does not mean you will always win by switching, but it is more likely. This is why I say when calculating probabilities you should take into account all available information.

8.2 Some Basic Ideas: Event Algebra

Probabilities and probability theory form the base of inferential testing, so if you want to do research you better get familiar with probability. Before getting into more technical aspects of calculating probabilities we need to go through a few basic ideas of where probability comes.
A **simple random experiment** is a procedure or operation performed where the outcome is unknown in advance, that is, prior to collecting data or performing any operation, the outcomes of the data collection

are unknown. The collection of all possible outcomes of a simple random experiment is the **sample space** that I will label ***S***. Here are few examples of simple random experiments and associated sample spaces:

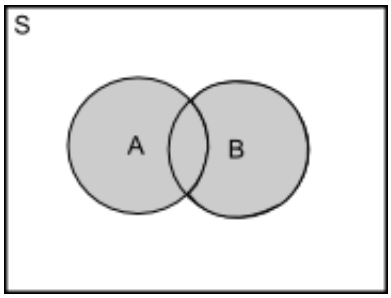
	Simple Random Experiment	***Sample Space***
1	Toss a coin exactly once	*S* = {H, T}
2	Toss a coin twice	S = {(H, H), (H, T), (T, H), (T, T)}
3	Rolling a six-sided die exactly once	S = {1, 2, 3, 4, 5, 6}
4	Rolling a six-sided die exactly twice	S = {(1, 2), (1, 2), (1, 3),...,(6, 5), (6, 6)}
5	Count the number of hairs on a cat	S = {1, 2, 3,...,*N*}, where *N* = total number of hairs
6	Select a card from a deck of playing cards	S = {Ace of Spaces...}

An **event** is a unique outcome in the sample space of a simple random experiment; it is a subset of the sample space. For example, when flipping a coin exactly once, the sample space includes two events that can occur: Heads and Tails. Here are some examples of events from the table above:

	Simple Random Experiment	***Event***
1	Toss a coin exactly once	Head = {H}
2	Toss a coin twice	Exactly one tail = {(H, T), (T, H)}
3	Rolling a six-sided die exactly once	Outcome is even = {2, 4, 5}
4	Rolling a six-sided die exactly twice	Sum is > 10 = {(5, 6), (6, 5), (6, 6)}
5	Count the number of hairs on a cat	Number of hairs is greater than 1000 = {1001, 1002, 1003,...,*N*}
6	Select a card from a deck of playing cards	Card is a King = {King of Hearts, King of Clubs, King of Diamonds, King of Spades}

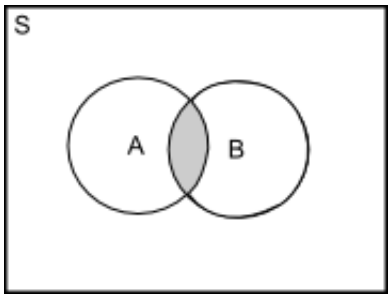
Event algebra is a method of expressing relationships among events and among combinations of events. It is a symbolic way to express relationships among events and among combinations of events from the sample space created in a random experiment. Below, I provide the symbolic representations of the more important relationships between events along with Venn diagrams to provide a graph representation

The **union** of two events, *A* and *B*, (symbolically A U B) is the event that consist of all outcomes that are associated with event A or event B or both. Thus, in a sample space *S*, you have events *A* and *B*; the union of those two events is the event that favors both events A and B or only event A or only event B. (See diagram at right, the shaded area represents all outcomes in the union of A and B). Here are some examples of the union of two events A and B:



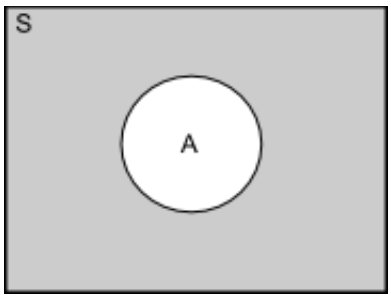
	A	***B***	***A U B***
1	Sum of two dice is multiple of two = {2, 4, 6, 8, 10, 12}	Sum of two dice is multiple of four = {4, 8, 12}	{2, 4, 6, 8, 10, 12}
2	Selected card is a two { two of spades, two of diamonds, two of clubs, two of hearts}	Selected card is a heart = {two of hearts, three of hearts,...ace of hearts}	{two of spades, two of diamonds, two of clubs, two of hearts, three of hearts,...,ace of hearts }

The **intersection** of two events, *A* and *B*, (symbolically A ∩ B) is the event that consist of all outcomes that are associated with event A and event B. Thus, in a sample space *S*, you have events *A* and *B*; the intersection of those two events is the event that favors both events A and B. (See diagram below)

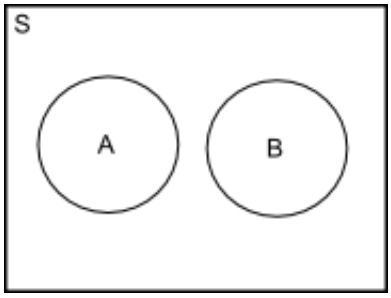


From Example 1 above, the intersection is A ∩ B = {4, 8, 12}, and the intersection from Example 2 is A ∩ B = {two of hearts}. That is, in Example 1, the values 4, 8, and 12 are the only outcomes that are both a multiple of two (event A) and a multiple of four (event B). For Example 2, the two of hearts is the only outcome that is a two (event A) and a heart (event B).

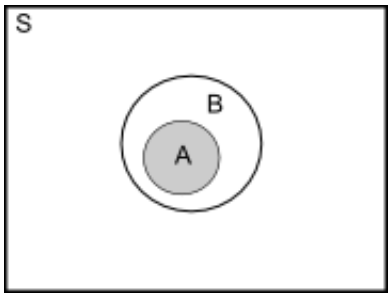
The **complement** of an event (symbolically A^C) is the event that consists of all outcomes that are not associated with event A. From Example 1, the complement of event A is A^C = {1, 3, 5, 7, 9, 11}, and the complement of event B is B^C = {1, 2, 3, 5, 6, 7, 9, 10, 11}. From example 2, the complement of event A is A^C = {all cards that are not two}, and the complement of event B is B^C = {all chards that are not hearts}.



If events A and B have no outcomes in common (symbolically A ∩ B = Φ, where Φ = “impossible”), then events A and B are **mutually exclusive** events. Thus, when two events are mutually exclusive the intersection of those two events is not possible. For example, if event A is selecting a biological male and event B is selecting a biological female then A ∩ B = Φ. As another example, if event A is select a club and even B is select a heart, then A ∩ B = Φ. In Examples 1 and 2 above, events A and B are not mutually exclusive, because the intersection of events A and B is possible in both examples.



An event is **included** in another events (symbolically A ⊂ B) if each and every outcome in event A is also in event B. This is similar to, but not the same as the intersection of A and B, because for the intersection of A and B, not all outcomes in A necessarily have to be in event B. For example, if event A is that the sum of two dice is a multiple of six and event B is the sum of two dice is a multiple of three, then event A must be included in event B (A ⊂ B), because all multiples of six will be multiples of three.



8.3 Probability Axioms and Things About Probability that are True

The probability of the outcomes in a simple random experiment is a function P that assigns a real number, P(E), to each event *E* in a sample space S, and satisfies the following three axioms:

1. $P(E) \geq 0$
2. $P(S) = 1$
3. $P(A \cup B) = P(A) + P(B)$ whenever events A and B are mutually exclusive

These three axioms form the basis of all probability theory. Simply, axiom 1 states the probability of an event must be greater than or equal to zero; axiom 2 states that the probability of observing *something* in the sample space is 1 (i.e., the probability that you will observe at least one event is 1); and axiom 3 states that the probability of observing at least one of two mutually exclusive events if the sum of their respective probabilities. The following are additional results that can be derived using event algebra and the three axioms of probability:

1. For any event, $P(A^C) = 1 - P(A)$
2. For any event, $P(A) \leq 1$
3. For all events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, whether A and B are mutually exclusive or not
4. For all events A and B, $P(A \cup B) = P(A \cap B) + P(A \cap B^C)$, whether A and B are mutually exclusive or not

To calculate the probability of an event (A), identify the number of possible outcomes favoring event A and divide that by the number of possible outcomes where event A *can* occur in the sample space (S):

$$P(A) = \frac{\text{number of outcomes in A}}{\text{number of outcomes in S}} = \frac{A}{S}$$

For example, we want to calculate the probability of selecting a Jack of Clubs from a deck of cards. In a standard deck of 52 cards there is only one jack of clubs, so A = 1 and the number of outcomes where event A *can* occur is S = 52, and the probability of selecting a jack of clubs is 1/52 = .019. Similarly, what is the probability of selecting any king, where event A = 'any king'. There are four possible kings in a deck of cards {king of clubs, king of spades, king of diamonds, king of hearts}, so A = 4 and S = 52 and the probability of selecting any king is 4/52 = .077. Thus, calculating a probability is as simple as dividing one value by another. The tricky part is figuring out what value should be divided by what; and this is where attention to detail is important.

8.4 Counting Rules: Permutations and Combination

It is often the case that several operations or procedures will be applied in a simple random experiment. Here is a simple rule to follow: If an operation, *O,* consists of a sequence of operations {O_1, followed by O_2,..., followed by O_r} and O_1 can result in any one of n_1 outcomes, O_2 can result in any one of n_2 outcomes

after performing O_1, and ... O_r can result in any one of n_r outcomes after performing O_1, O_2,..., O_{r-1}, then the operation O can result in any one of $n_1 n_2 ... n_r$ outcomes. Basically, if an operation (simple random experiment) consists of many operations and each operation can result in one of several possible outcomes, then for the simple random experiment any one of a number of outcomes will result. Sounds complicated right? It's really not:

For example, in the behavioral sciences it is often the case that researchers have a certain number of conditions that can be presented to subjects and the researcher may want to know all possible **permutations** that can be presented to subjects. That is, how many orders in which the conditions can be presented. A permutation is an ordered arrangement of distinct items. The total number of permutation of *n* distinct items is:

$$n!$$

That is, "*n* factorial." The factorial (!) operator, tells you to multiply the number (*n*) by all numbers less than *n* not including zero. That is, 4! is shorthand for writing 4 X 3 X 2 X 1 = 24. Thus, whenever you see this symbol you should expand the expression preceding it and multiply the given number by all values below it, but not zero. Remember 1! = 1 and 0! = 0. There is a factorial table in APpendix A (Table 7). For example:

2! = 2 x 1 = 2
3! = 3 x 2 x 1 = 6
4! = 4 x 3 x 2 x 1 = 24

Using a more concrete example, say an investigator is conducting a taste-preference survey for different types of scotch.. The investigator has three types of scotch that can be presented and wants to know all of the permutations that are possible for the three scotches:

Permutation	Scotch 1	Scotch 2	Scotch 3
1	A	B	C
2	A	C	B
3	B	A	C
4	B	C	A
5	C	A	B
6	C	B	A

However, the researcher may want to present only a subset (*r*) of all the possible conditions or outcomes. In this case the researcher may want to know how many permutations are possible from all of the possible subsets out of the total number of conditions. The number of permutations of *r* items out of *n* items is:

$$_{n}P_{r} = \frac{n!}{(n-r)!}$$

For example, the number of different permutations for r = 3 conditions out of n = 10 total conditions is:

$$_{10}P_{3} = \frac{10!}{(10-3)!} = \frac{10\times9\times8\times7\times6\times5\times4\times3\times2\times1}{7\times6\times5\times4\times3\times2\times1} = \frac{6328800}{(5040)} = 760$$

Using the scotch example above, the researcher might want to know the number of permutations of two scotches out of the three scotches:

$$_{3}P_{2} = \frac{3!}{(3-2)!} = \frac{3\times2\times1}{1} = \frac{6}{1} = 6$$

Alternatively, a researcher may simply want to know how many possible **combinations** of subsets are possible from some larger number of conditions. While permutations are the total number of possible orders

in which a set of conditions can be presented, combinations are the total number of sets that can be created from some larger number of conditions, but where order of conditions is not important.

Say we have n = 10 conditions and we want to determine the number of combinations and permutations possible for subsets of r = 3 conditions. That is, we have 10 conditions that we can present to subjects, but we only want to present 3 to each subject; so we want to get an idea of how many subjects are needed. The formula for calculating the number of combinations is:

$$ {}_{n}\mathrm{C}_{r} = \frac{n!}{(n-r)!r!} $$

From the example above, the number of different combinations of r = 3 conditions out of n = 10 conditions:

$$ {}_{10}C_{3} = \frac{10!}{(10-3)!3!} = \frac{10\times9\times8\times7\times6\times5\times4\times3\times2\times1}{(7\times6\times5\times4\times3\times2\times1)(3\times2\times1)} = \frac{6328800}{(5040)(6)} = \frac{6328800}{30240} = 120 $$

Using the scotch example, assume that the researcher wants to know the number of combinations of two scotches out of three scotches:

$$ {}_{3}C_{2} = \frac{3!}{(3-1)!2!} = \frac{3\times2\times1}{(1)2\times1} = \frac{6}{2} = 3 $$

We'll now explore calculating probabilities using the ideas of permutations and combinations from above.

Example 1: What is the probability of getting exactly one club and exactly one jack in two randomly dealt cards; assuming order does not matter?

Event A = one club and one jack. There are 13 clubs and 13 jacks, and the number of unordered combinations of clubs and jacks is 13 x 13 = 169

Sample Space, S = all unordered two card hands. Thus, the sample space is all combinations of two cards out of a deck of 52 cards (I have simplified the math to the right):

$$ {}_{52}\mathrm{C}_{2} = \frac{52!}{(52-2)!2!} $$

$$ {}_{52}\mathrm{C}_{2} = \frac{52\times51}{2} $$

$$ {}_{52}\mathrm{C}_{2} = 1326 $$

Thus, P(A) = 169/1326 = 0.127

Example 2: What is the probability that a five-card poker hand contains exactly three clubs?

Event A = three clubs and two non-clubs. In this case, you need to account for the combinations of three clubs of the 13 total clubs and also account for the combinations of two cards out of the 39 non-clubs. Thus:

For the clubs:

$$ {}_{13}\mathrm{C}_{3} = \frac{13!}{(13-3)3!} $$

$$ {}_{13}\mathrm{C}_{3} = \frac{13\times12\times11}{3\times2\times1} $$

$$ {}_{13}\mathrm{C}_{3} = 286 $$

For the non-clubs:

$$ {}_{39}\mathrm{C}_{2} = \frac{39!}{(39-2)2!} $$

$$ {}_{39}\mathrm{C}_{2} = \frac{39\times38}{2\times1} $$

$$ {}_{39}\mathrm{C}_{2} = 741 $$

Sample Space, S = all five card hands Thus, the sample space is all combinations of five cards out of a deck of 52 cards (I have simplified the math):

$$ {}_{52}\mathrm{C}_{5} = \frac{52!}{(52-5)!5!} $$

$$ {}_{52}\mathrm{C}_{5} = \frac{52\times51\times50\times49\times48}{5\times4\times3\times2\times1} $$

$$ {}_{52}\mathrm{C}_{5} = 2598960 $$

Thus, P(A) = (286 * 741)/2598960 = 0.082

Example 3: What is the probability of getting five cards from the same suit?

Event A = all five cards are from the same suit. In this case, we are interested in the number of five card combinations from each of four 13-card suits. The number of five card combinations from each suits:

$$_{13}C_5 = \frac{13!}{(13-5)!5!} = \frac{13\times12\times11\times10\times9}{5\times4\times3\times2\times1} = 1287$$

There are four suits, so the number of five card combinations from the four suits is: 4 x 1287 = 5148. That is, there are 1287 five-card combinations of hearts, diamonds, spades, and clubs (1287 + 1287 +1287 +1287 = 5148).

Sample Space, S = all unordered five card hands Thus, the sample space is all combinations of five cards out of a deck of 52 cards (I have simplified the math):

$$_{52}C_5 = \frac{52!}{(52-5)!5!}$$

$$_{52}C_5 = \frac{52\times51\times50\times49\times48}{5\times4\times3\times2\times1}$$

$$_{52}C_5 = 2598960$$

Thus, P(A) = 5148/2598960 = 0.002

8.5 Probability and Sampling from a Population

One issue with calculating probability is how selecting random samples from a population can change the probability of future selections. As discussed earlier, when selecting a sample you typically engage in random sampling, which literally means selecting cases from a population at random and without bias. Random sampling is good, because such a procedure should result in a sample that contains the individual differences that are present in the population. If so, you have an unbiased random sample.

One issue is whether you **sample *with* replacement** or **sample *without* replacement**. Sampling *with* replacement occurs when after a case is selected it is returned (replaced) to the population. Sampling *without* replacement occurs when after a case is selected it is not returned to the population. There are pros and cons of each method.

When sampling *with* replacement the population size (*N*) never changes, so the probability of an event never changes. But, because you are returning cases to the population you run the risk of selecting the same case more than once. Sampling *without* replacement corrects for the possibility of selecting a case multiple times, because after a case is selected it is not returned to the population. But, because after each selection the population size (*N*) will decrease one, the probability of unselected events being selected increases over time. Because most population are large (on the order of millions), altering the number of cases that can be selected should not make much of a difference; thus, most behavioral scientists use sampling without replacement.

As examples of sampling with and without replacement, I'll use the statistics class example above with *N* = 50 students. This class includes 25 psychology majors, 10 neuroscience majors, 10 communications majors and 5 counseling majors. Say we select a student at random from the class. The initial probabilities of selecting a student from each major are listed in the table below:

Major	***f*(Major)/N**	**p(Major)**
Psychology	25/50	.500

Neuroscience	10/50	.200
Communications	10/50	.200
Counseling	5/50	.100

After selecting a student (regardless of major), the student was returned to the class (sampling with replacement), and *N* will still be 50. Let's say that say we select a Communications major, but the student was not returned to the class (sampling *without* replacement). Because *N* now equals 49 the probability that students from the other majors will be selected changes, which can be seen in the table below. Notice the probability of selecting each major increased, except for that of Communications majors, because that major was decreased by one student:

Major	***f*(Major)/N**	**p(Major)**
Psychology	25/49	.510
Neuroscience	10/49	.204
Communications	9/49	.184
Counseling	5/49	.102

Say that we select a second student, a psychology major, and do not return that student to the class. Because *N* now equals 48 the new probabilities for a student from each major being selected can be seen in the table below. The point is that changing the underlying population size, just like changing information in a situation (e.g., Monty Hall problem), alter probabilities of selecting events:

Major	***f*(Major)/N**	**p(Major)**
Psychology	22/48	.500
Neuroscience	10/48	.208
Communications	9/48	.188
Counseling	5/48	.104

8.6 Probabilities Based on Multiple Variables

Assume a researcher is interested in the association between political orientation and sexual behavior. The researcher surveys all incoming freshmen at a university (*N* = 1000 freshmen) on two variables: (1) political attitude {liberal and conservative} and (2) whether the student has engaged in intercourse {yes and no}. Thus, the freshmen can be classified into one of four groups: (1) liberals who have had intercourse, (2) liberals who have not had intercourse, (3) conservatives who have had intercourse, and (4) conservatives who have not had intercourse. For argument's sake, assume all freshmen are 18 years old and that there are equal numbers of males and females in each group.

When you have a combination of variables where frequencies are measured you have a **contingency table** of that data (see below). A contingency table is used to record the relationship between two or more variables. Thus, to examine the relationship between political attitude and sexual behavior the frequencies of students in each of those four groups will be examined in a tabled form, like that below:

	Political Attitude		
Previously Engaged in Intercourse?	**Liberal**	**Conservative**	**Totals**
Yes	500	100	600
No	200	200	400
Totals	700	300	*N* = 1000

Each of the four groups created by combining the two variables mentioned above is referred to as a **cell**. The value in each cell are the frequencies of students belonging to that combination of variables; thus, 500 liberal freshmen reported having had intercourse, 100 conservative freshmen reported having had intercourse, and 200 liberal freshmen as well as 200 conservative freshmen reported not having had

intercourse. Values in the last column and the bottom row (*Totals*) are the sum of the cell frequencies for that row or column; and these are the **marginal frequencies**.

Using this contingency table probabilities of events, combinations of two events, and probabilities that are conditional on (dependent on) some other event can be calculated. When going through each of the following probabilities, keep in mind that for each of the following probabilities, you must identify the number of observations favoring the event of interest ('event A'), and the total number of *relevant* observations where that event can occur

8.6.1 Simple Probabilities Say the researcher wants to determine the probability of selecting a liberal student from among all incoming freshmen. He needs to identify the number of liberal students and the total number of students that *could* be a liberal student, which is the total frequency (N). If we call selecting a liberal student event A, then the number of liberal students is A = 700 and the total number of students that could have been a liberal student is the total number of freshmen, *N* = 1000. Thus, the probability of selecting a liberal is:

$$\mathrm{P}(\mathrm{Liberal}) = \frac{\text{number of liberals}}{\mathrm{N}} = \frac{700}{1000} = 0.7$$

Likewise, the probability of selecting a conservative student is the total number of conservative students (*f* = 300) divided by the total number of students (*N* = 1000):

$$\mathrm{P}(\mathrm{Conservative}) = \frac{300}{1000} = 0.3$$

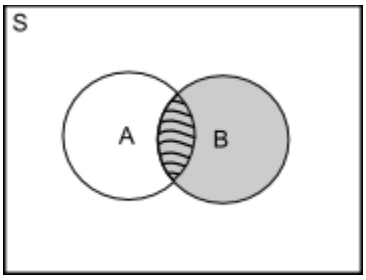
These are **simple probabilities**, which are the probability of selecting a single event or level of a variable. That is, in both cases, the researcher calculated the probability of one political orientation without taking into account the variable *'previously engaged in intercourse?'*

Notice that the two probabilities add to 1.0. When this occurs it means that the outcomes for this variable (political attitude) are **exhaustive**. That is, all of the possible outcomes of the variable have selected or identified (the researcher has determined the probabilities for all possible events for the variable political attitude). Also, because a student can either be a liberal or can be a conservative, but cannot be a political conservative and a political liberal; this means that the two outcomes for Political Attitude are **mutually exclusive**. That the levels of a variable are mutually exclusive means that an case (any freshman student) can belong to only one level of a variable, but cannot belong to more than one level. The variable *previously engaged in intercourse* is also mutually exclusive, because a student can have previously engaged in sexual intercourse, or not; that is, a student cannot have both engaged in and not engaged in intercourse.

8.6.2 Conditional Probabilities Conditional probabilities can be used to assess whether there is a relationship between the variables. A conditional probability is the probability of selecting some event (A) given some other event (B) is selected, that is, that some other condition is satisfied. A conditional probability is written as:

$$\mathrm{P}(\mathrm{A}\,|\,\mathrm{B}) = \frac{\text{number of outcomes favoring A and B}}{\text{number of outcomes favoring}} = \frac{\mathrm{A} \cap \mathrm{B}}{\mathrm{B}}$$

P(*A*|*B*) is read *the probability of event A <u>given</u> event B*. In the Venn diagram to the right, the hatched area is the conditional probability of selecting A given you have selected B, and the B area is shaded to signify that we are interested in outcomes only in event B.



From the contingency table above, say the researcher wants to calculate the probability of selecting a student that has engaged in intercourse given the student is also a political conservative. Stated differently, *if a selected student is a conservative what is the probability s/he has engaged in intercourse?* In this case, event A is having engaged in intercourse and event B is students who identify themselves as conservative. Why is this so? This question is asking the probability of selecting a conservative student who has had intercourse, out of only those students who report being politically conservative. Thus, event B (the denominator above) needs to be the number of cases that *can* be a conservative student who has engaged in intercourse, which is all of the conservative students. Event A, the event we are slightly more interested in, then, must be students who have reported having intercourse.

From the contingency table, the number of conservative students is B = 300. This value comes from the marginal frequencies. Out of these 300 conservative students, 100 reported to have engaged in sexual intercourse; that is, A ∩ B = 100. This value comes from the cell frequencies. Thus, the conditional probability that a student has engaged in intercourse *given* the student is a conservative:

$$P(\text{Yes} \mid \text{Conservative}) = \frac{\text{Yes} \cap \text{Conservative}}{\text{Conservative}}$$

$$P(\text{Yes} \mid \text{Conservative}) = \frac{100}{300} = 0.333$$

There is about at 33.3 percent chance of selecting a student who has had sexual intercourse, if you have also selected a conservative student. As another example, let's say that the researcher is now interested in the probability of selecting a liberal student *given* we have selected a student who has not had intercourse. In this example, event A would be 'liberal students', and event B would be all students who have reported not having had intercourse. Thus, the number of liberal freshmen students who have not had intercourse is A ∩ B = 200, and the total number of students who have not had intercourse if B = 400. The probability of selecting a liberal student given we have already selected a student who has not engaged in intercourse is:

$$P(\text{Liberal} \mid \text{No}) = \frac{200}{400} = 0.5$$

Thus, there is about a 50 percent chance that the researcher has selected a liberal freshmen student if he has selected a student who has not engaged in intercourse.

Conditional probabilities can be used to determine whether there is a relationship between the two events (and variables), that is, whether the two events are **independent** or not. If two events are independent it means that selecting one event does not depend on the other event. In contrast, if two events are not independent, that is, there is a relationship between the events then selecting one event does depend on the other. Two events are *independent* if the following holds: **P(A) = p(A|B)**

That is, if the probability of event A is equal to the conditional probability of event A given event B, then the two events A and B are independent and selecting event A does not depend on selecting B. Two events are not independent, and there is a relationship between them, if the following holds: **P(A) ≠ p(A|B)**

To determine whether having previously engaged in intercourse (event A) is independent of being a conservative (event B), the first conditional probability that was calculated above must be compared to the probability of having engaging in intercourse, which is:

$$P(\text{Yes}) = \frac{\text{Yes}}{N} = \frac{600}{1000} = 0.6$$

From above, we know that,

$$P(\text{Yes} \mid \text{Conservative}) = \frac{100}{300} = 0.333$$

Because *P*(*Yes*) ≠ *P*(*Yes* | C*onservative*) some relationship must exist between having engaged in intercourse and being a conservative; that is, having had intercourse is influenced by being a conservative (versus being a liberal). Thus, selecting a freshmen student who has engaged in sexual intercourse does depend on selecting a politically conservative freshmen student.

8.6.3 Joint Probabilities If you wish to determine the probability of a combination of two events out of all events you must calculate a **joint probability**. A joint probability is simply the probability of selecting tow events, A and B, together. A joint probability is conceptualized as:

$$P(A \cap B) = \frac{A \cap B}{N}$$

Note that *P*(*A* ∩ *B*) is read *the probability of event A and event B*. From the contingency table above, say that the researcher wants to calculate the probability of selecting a student that has not engaged in intercourse (event A) *and* is a politically liberal student (event B). The number of students who have not engaged in intercourse *and* are liberals is A ∩ B = 200 and N = 1000, thus, the joint probability of selecting a student that has not had intercourse and is a liberal is:

$$P(\text{No} \cap \text{Liberal}) = \frac{\text{No} \cap \text{Liberal}}{N}$$

$$P(\text{No} \cap \text{Liberal}) = \frac{200}{1000} = 0.2$$

A joint probability is a cell frequency divided by the total frequency (N). This provides you with the probability of selecting combination of events, thus, the probability of selecting a student who has not had intercourse and is a politically liberal student is p = 0.2.

8.6.4 Addition Rule of Probability The addition rule of probability is used to determine the probability of *at least* one of two different events, *or both* events, occurring. That is, what is the probability of only event A, only event B, or both events A and B occurring? The addition rule of probability is written as:

$$\mathbf{P(\mathit{A} \cup \mathit{B}) = \mathit{p}(\mathit{A}) + \mathit{p}(\mathit{B}) - \mathit{p}(\mathit{A} \cap \mathit{B})}$$

Thus, to find the probability of event A or event B (or both) occurring, you add the simple probability of event A to the simple probability of event B, and subtract the joint probability of both events A and B occurring. Note that *p* (A ∩ B) will be zero if the two events, A and B, are mutually exclusive.

Say that the researcher wants to calculate the probability of selecting a student who has engaged in intercourse (event A) or is politically liberal (event B), or had intercourse and is liberal. The probability of selecting a student who has engaged in intercourse is 600/1000 = 0.6. The probability of selecting a student who is a political liberal is 700/1000 = 0.7. The joint probability of selecting a student who has had intercourse *and* is a political liberal is 500/1000 = 0.5. Thus, the probability of selecting a student who has engaged in intercourse or a liberal student is:

P(*Had Intercourse* U *Liberal*) = 0.6 + 0.7 - 0.5 = 0.8

Thus, there is an 80 percent chance of selecting at least a student how has had intercourse or is a politically liberal student.

You subtract the joint probability of events A and B because the two events A and B may not be mutually exclusive. If events A and B are not mutually exclusive, which is the case here because you can have liberal students who have had intercourse, then the sum of the probabilities of events A and B might be greater than 1, which would also be the case here. Remember, probabilities can range from between 0 to 1 only. To correct for this you must factor out the joint probability of both events. But again, if events A and B are mutually exclusive, then $p(A \cap B)$ will be equal to 0, because being mutually exclusive means that events A and B do not occur together; hence, the frequency of events A and B occurring together will be 0.

8.6.5 Multiplication Rules of Probability The multiplication rule is used to calculate the probability of selecting two events, A and B, like the joint probability. In some ways the multiplication rule is just another manifestation of finding joint probabilities, but can also be used to express relationships or lack thereof between events. There are two ways that the multiplication rule can be applied: (1) when events are independent, and (2) when events are not to be independent. If two events, A and B, are independent and there is no relationship between the events, such that $P(A) = P(A|B)$, then the following multiplication rule can be applied to events A and B, to determine the probability of observing both events:

$$P(A \cap B) = p(A)p(B)$$

Thus, if two events, A and B, have no known relationship between them; to find the probability of both events occurring [p(A , B)], simply multiply the simple probability of event A by the simple probability of event B. As an example, say that we want to know the probability of rolling a '2' on a six-sided die (event A) and flipping a 'tails' on a two sided coin (event B). The joint probability of these two events can be found by applying the multiplication rule for independent events:

$$P(2 \cap tails) = P(2)P(tails)$$
$$P(2 \cap tails) = (1/6)(1/2)$$
$$P(2 \cap tails) = (0.167)(0.5)$$
$$P(2 \cap tails) = .084$$

As another example, what is the probability of flipping tails on a two-sided coin (event A) and then flipping heads on the same coin (event B). If the outcome of the second flip does not depend on the first flip we can answer this by applying the multiplication rule for independent events:

$$P(tails \cap heads) = P(tails)P(heads)$$
$$P(tails \cap heads) = (1/2)(1/2)$$
$$P(tails \cap heads) = (.5)(.5)$$
$$P(tails \cap heads) = .25$$

Using the contingency table from above, what is the probability that the researcher would select a politically conservative student, return that student to the population, and select another politically conservative student? Because the researcher sampled with replacement, N does not change, hence we can answer this by applying the multiplication rule for independent events:

$$P(conservative \cap conservative) = P(conservative)P(conservative)$$
$$P(conservative \cap conservative) =(300/1000)(300/1000)$$
$$P(conservative \cap conservative) =(0.3)(0.3)$$
$$P(conservative \cap conservative) =0.09$$

Note that this works only when sampling with replacement. If sampling without replacement, then the probability of selecting the second conservative student *does* depend on selecting the first student. In this case, the two events are not independent and you would have to apply the multiplication rule for non-independent events, which we discus next.

If two events A and B are not independent and the events have some relationship, such that $P(A) \neq P(A \mid B)$, then the following multiplication rule can be applied to events A and B, to determine the probability of observing both events:

$$\mathbf{P(A \cap B) = P(B)P(A|B)}$$

The probability of observing two non-independent events A and B is equal to the probability of event B multiplied by the conditional probability of event A given event B. Also, note that:

$$\mathbf{P(A \cap B) = P(A)P(B|A)}$$

Thus,

$$\mathbf{P(B)P(A|B) = P(A)P(B|A)}$$

From the contingency table, say the researcher wants to calculate the probability of selecting a student who had intercourse (event A) who is politically conservative (event B). Essentially, this is asking for the joint probability of selecting freshmen student who has intercourse and is a conservative. This question is, of course easily answered by dividing the number of conservative freshmen who have not had intercourse (100) by the total frequency (1000); thus, p(Yes , Conservative) = 100/1000 = 0.1. But the multiplication rule for non-independent events is another way to answer the question.

From the contingency table, the probability of selecting a conservative freshmen student is P(Conservative) 300/1000 = 0.3 and the conditional probability of having had intercourse (event A) given the student is conservative (event B) is P(Yes | Conservative) = 100/300 = 0.333. Using these two probabilities and the multiplication rule we can determine the probability of selecting a freshmen student who has not had intercourse and is a Conservative:

$$\mathbf{P(Yes \cap Conservative) = P(Conservative)P(Yes \mid Conservative)}$$
$$\mathbf{P(Yes \cap Conservative) = (0.3)(0.333)}$$
$$\mathbf{P(Yes \cap Conservative) = 0.1}$$

Note that this is equal to the joint probability calculated above; that is, P(Yes ∩ Conservative) = 100/1000 = 0.1.

Now say that the researcher is interested in the probability of selecting a politically conservative student, not returning that student to the population, and then selecting another politically conservative student? Because the researcher sampled *without* replacement, N does change and the probability of selecting the second conservative student depends on selecting the first conservative student. We can answer this by applying the multiplication rule for non-independent events. But first, here is the contingency table prior to selecting anyone:

	Political Attitude		
Previously Engaged in Intercourse?	**Liberal**	**Conservative**	**Totals**
Yes	500	100	600
No	200	200	400
Totals	700	300	*N* = 1000

If the researcher selects a conservative student and does not replace that student to population, the resulting contingency table would look like this:

	Political Attitude		
Previously Engaged in Intercourse?	**Liberal**	**Conservative**	**Totals**

Yes	500	?? (99 or 100)	600
No	200	?? (199 or 200)	400
Totals	700	299	*N* = 999

Notice, the number of conservative students is now 299. Since we do not know whether the conservative student who was selected had intercourse or not, I have simply indicated that there may be the original number of those students or one less. But that point is not important here.

P(conservative 1 ∩ conservative 2) = P(conservative 2)P(conservative 2 | conservative 1)
P(conservative 1 ∩ conservative 2) = (300/1000)(299/999)
P(conservative 1 ∩ conservative 2) = (0.3)(0.299)
P(conservative 1 ∩ conservative 2) = 0.0987

Note that this does round to 0.09, but I wanted to show that there is a difference between sampling with and sampling without replacement.

8.6 Bayes' Theorem

From the preceding section, the multiplication rule for non-independent events can be algebraically rearranged to solve for the conditional probability of event A given event B, using the joint probability of events A and B divided by the probability of event B. Thus,

$$P(A \cap B) = P(B)P(A \mid B)$$

This can be rearranged into:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Thus, the conditional probability of selecting event A given you have selected event B, is equal to the joint probability of selecting both events A and B divided by the probability of selecting only event B. In cases where you do not know P(A ∩ B), you can apply the multiplication rule for non-independent events. That is, we know that:

$$P(A \cap B) = P(A)P(B \mid A)$$

By plugging this produce into the numerator from our rearranged expression above, we get:

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

This expression is one form of **Bayes' Theorem**, which is used for computing the conditional probability of an event A given event B. It is another way of calculating a conditional probability of an event when you know something about the other events.

For example, say that we want to calculate the probability of selecting a student who has had intercourse (event A) given they are a conservative student (event B); thus, we want to know p(Yes | Conservative), which from earlier was 100/300 = 0.333. We can also find this probability using Bayes' Theorem:

$$P(\text{Yes} \mid \text{Conservative}) = \frac{P(\text{Yes})P(\text{Conservative} \mid \text{Yes})}{P(\text{Conservative})}$$

I have reproduced the contingency table below, so it will be easier to see where the various probabilities are coming from.

	Political Attitude		
Previously Engaged in Intercourse?	**Liberal**	**Conservative**	**Totals**
Yes	500	100	600
No	200	200	400
Totals	700	300	*N* = 1000

First, we need to know the probability of selecting a student who has had intercourse, which is *p*(Yes) = 600/1000 = 0.6. Next, we need to know the conditional probability of selecting a conservative student given they have engaged in intercourse, which is p(Conservative | Yes) = 100/600 = 0.167. Finally, we need the probability of selecting a conservative student, which is p(Conservative) = 300/1000 = 0.3. Plugging these values into Bayes' Theorem, we get:

$$P(\text{Yes}\,|\,\text{Conservative}) = \frac{P(\text{Yes})P(\text{Conservative}\,|\,\text{Yes})}{P(\text{Conservative})}$$

$$P(\text{Yes}\,|\,\text{Conservative}) = \frac{(0.6)(0.167)}{(0.3)}$$

$$P(\text{Yes}\,|\,\text{Conservative}) = \frac{0.1}{(0.3)}$$

$$P(\text{Yes}\,|\,\text{Conservative}) = 0.333$$

This was the conditional probability calculated earlier. The important point to remember is that you can use relationships between variables and events to calculate probabilities of observing specific outcomes from the combinations of events. Consequently, you can use the simple, conditional, and joint probabilities in such ways to determine if relationships between variables exist. Indeed, this is just one of the many uses of Bayes' Theorem, which we will likely cover in more detail during lecture.

8.7 Back to Monty Hall

Now that you know a little about probability and understand the most important thing when calculating a probability is to take information into account, let's go back to the Monty Hall problem. Remember, after you make your initial selection and Monty Hall opens a door exposing a goat the probability of winning by staying with your original choice of door is 1/3 and the probability of winning through switching to the other unopened door is 2/3. Why...

The table below displays the Monty Hall problem. That is, there are three doors: Door 1, Door 3, and Door 3. A prize is hidden behind one door (assume Door 1) and goats are hidden behind the other two doors. The probability that you choose any one of the three doors is 1/3, thus the probability of winning on your first selection is 1/3. The probabilities change only *after* Monty Hall opens a door.

You Pick ↓	**Door 1**	**Door 2**	**Door 3**	**Result of Switch?**	**Result of Stay?**	**p(Scenario)**
Door 1	You	MH opens		Lose	Win	1/6
Door 1	You		MH opens	Lose	Win	1/6
Door 2		You	MH opens	Win	Lose	1/3
Door 3		MH opens	You	Win	Lose	1/3

Say you initially choose Door 1, which is the door that hides the prize. In this case Monty Hall can open Door 2 or Door 3 to expose a goat and the probability that Monty Hall opens either door is 1/2. Hence, from the 1/3 chance you select Door 1, there is 1/6 chance Monty Hall will open Door 2 and a 1/6 chance he will open Door 3. This 1/6 comes from (1/3)(1/2) = 1/6. In either case, if you *stay* with Door 1 (the door with the prize) you win if you stay and you lose if you switch. Thus, the probability of winning by staying with you first choice is 1/3, if you have indeed selected the correct door on the first try.

On the other hand, say you initially pick Door 2, which hides a goat. In this case, because the initial pick was a door with a goat Monty Hall is forced to open the door that hides the other goat. Remember, Monty Hall knows which door hides the prize! Thus his opening a door gives you information about where the prize might be hidden. From the 1/3 chance you select Door 2, the probability that Monty Hall will open Door 3 to expose the other goat 1.0. This is because, if you have selected a door with a goat, Monty Hall cannot open up the prize door and he cannot open your door, he must open up the only other door with a goat. In this case if you switch you win, but if you stay you lose.

Finally, say that you initially pick Door 3, which also hides a goat. Again, because your initial pick was a door with a goat Monty Hall is forced to open Door 2, which hides the other goat. From the 1/3 chance you select Door 3, the probability that Monty Hall opens Door 2 is 1.0. In this case if you switch you win, but if you stay you lose.

In the end, there is a 1/3 chance that you correctly select the door with the prize as your first choice. In this case, if you stay you win and if you switch you lose. There is a 2/3 chance that you will initially select a door with a goat. In this case, if you stay you lose and if you switch you win. Thus, there is a 2/3 chance of winning by switching and a 1/3 chance of winning by staying. This, in this non-intuitive scenario, it is beneficial to switch your initial choice, because this maximizes your chance of winning the prize. Indeed, if you look at the statistics from the game show “Let’s Make a Deal”, people won after a switch about 2/3 of the time, and people won from staying about 1/3 of the time.

The point is that information about events, whether subtle as in the case of Monty Hall or more overt, provides information that can seriously alter a probability structure; and this information should always be taken into account.

CH 8 Homework Questions

1. What is the difference between sampling with replacement versus sampling without replacement?

2. For each of the following, assume you are selecting cards from a standard deck of 52 playing cards.
   a. What is the probability that a randomly selected card is a jack?
   b. What is the probability that a randomly selected card is a king or a jack?
   c. What is the probability that a randomly selected card is a four or an ace?
   d. What is the probability that a randomly selected card is an eight or a heart?
   e. What is the probability that a randomly selected card is a seven and a queen?
   f. What is the probability that a randomly selected card is a diamond and a six?
   g. What is the probability that a randomly selected card is a diamond, a king, or a spade?

3. *Use this information to answer the questions that follow*: Jack has a bag of magic beans. His bag has contains 5 red beans, 2 blue beans, 10 green beans, 7 black beans, and 6 sparkled beans. Based on this information, answer the following questions, and be sure to incorporate jack's prior actions selections into your answer for each successive question.
   a. Jack reaches into his bag, selects a bean, looks at it, and replaces it. What is the probability this was a blue bean?
   b. Jack reaches back into his bag, selects a bean, looks at it, and eats it. What is the probability this was a red bean?

c. Assume that the bean that jack selected in b was green. Jack reaches back into his bag, selects a bean, looks at it, and then replaces it. What is the probability that jack selected a black bean?
d. Jack reaches back into his bag, selects a bean, looks at it, and eats it. What is the probability this was a green bean?
e. Assume the bean Jack selected in d was green. Jack reaches into his bag and pulls out a bean, looks at it, and eats it. What is the probability that this bean was a sparkled bean or a black bean?

Use this information to answer Exercises 4 – 18: A political scientist interviewed 500 professors at university to study the relationship between area of scholarship and attitudes toward embryonic stem cell research. The following contingency table was observed:

	Attitude toward Stem Cell Research			
Area of Scholarship	Favors	Opposes	No Opinion	Totals:
Natural Sciences	90	20	10	120
Behavioral Sciences	60	20	30	110
Theology	30	110	10	150
Humanities	20	50	50	120
Totals:	200	200	100	$n_T = 500$

4. Are levels of the independent variable 'Attitude toward stem cell research' mutually exclusive?

5. What is the probability that a randomly selected individual favors stem cell research?

6. What is the probability that a randomly selected individual has no opinion on stem cell research?

7. What is the probability that a randomly selected individual studies the natural sciences?

8. What is the probability than a randomly selected individual studies theology?

9. What is the probability than a randomly selected individual studies the behavioral sciences and has no opinion on stem cell research?
10. What is the probability that a randomly selected individual favors stem cell research given the individual is a theologian?

11. What is the probability that a randomly selected individual is a theologian given the individual favors stem cell research?

12. Why are the probabilities in 10 and 11 different?

13. Are being a natural scientist and opposing stem cell research independent? Why or why not?

14. What is the probability that a randomly selected individual studies the behavioral sciences and also favors stem cell research?

15. What is the probability that a randomly selected individual is a theologian and opposes stem cell research?

16. What is the probability that a randomly selected individual studies in the humanities or favors stem cell research?

17. What is the probability that a randomly selected individual is a behavioral scientist or has no opinion on stem cell research?

18. What is the probability that a randomly selected individual favors stem cell research or opposes stem cell research?

Use the following information to complete Exercises 19 – 32: A political scientist interviewed 500 people to study the relationship between political party identification and attitudes toward public sector unionization. The following contingency table was observed:

	Attitude Toward Public Sector Unionization		
Political Party identification	Favors	Opposes	Totals
Democrat	160	40	200
Republican	40	160	200
Independent	90	10	100
Totals	290	210	N = 500

19. What is the probability than an individual favors public sector unionization?

20. What is the probability than an individual opposes public sector unionization?

21. What is the probability than an individual is a Democrat?

22. What is the probability than an individual is a Republican?

23. What is the probability than an individual is an Independent?

24. What is the probability that an individual favors public sector unionization given the individual is a Democrat?

25. What is the probability that an individual is a Democrat given that the individual favors public sector unionization?

26. Are being a Democrat and opposing public sector unionization independent? Why or why not?

27. What is the probability that an individual is a Republican who favors public sector unionization?

28. What is the probability that an individual is a Republican who opposes public sector unionization?

29. What is the probability that an individual opposes public sector unionization given they are a Republican?

30. What is the probability that an individual is an Independent who opposes public sector unionization?

31. What is the probability that an individual is a Republican or favors public sector unionization?

32. What is the probability that an individual is an Independent or opposes public sector unionization?

Use the following information to complete Exercises 33 – 41: I asked n = 1000 students the following questions: (a) are you an underclassman (freshman/sophomore) or are you an upperclassman (junior/senior); and (b) do you own the latest Tim McGraw album. Use this contingency table to answer the following questions:

	"Own the Latest Tim McGraw Album?"	
College Year	Yes	No
Underclassman	300	200
Upperclassman	400	100

33. Is the variable “Own the Latest Tim McGraw album” mutually exclusive? Why or why not?

34. What is the probability that a randomly selected student owns the Tim McGraw album?

35. What is the probability that a randomly selected student is an underclassmen?

36. What is the probability that a randomly selected student owns the Tim McGraw album given that he/she is an upperclassman?

37. What is the probability that a randomly selected student is an underclassman given that he/she owns the Tim McGraw album?

38. Is the probability that a randomly selected student owns the Tim McGraw album independent of being an upperclassman? Why or why not?

39. What is the probability that a randomly selected student owns the Tim McGraw album and is an upperclassman?

40. What is the probability that a randomly selected student is an underclassman, or owns the Tim McGraw album, or both?

41. What is the probability that a randomly selected student is an underclassman or is an upperclassman?

42. Two events, A and B, are mutually exclusive. The probability of event A is 0.200 and the probability of event B is 0.500, what is the probability of at least one of the two events occurring?

43. The probability of event B is 0.410, and the probability of event A given event B is 0.850. What is the probability of both events A and B occurring?

44. Two events, A and B, are independent. The probability of event A is 0.35 and the probability of event B is 0.480. What is the probability of both event A and event B occurring?

45. What is the probability of flipping a heads on an unbiased, two-sided coin and then rolling a ‘6’ on an unbiased, six-sided die?

46. What is the probability of flipping a heads on an unbiased, two-sided coin three times in a row?

47. The probability of some event A is 0.800, the probability of some other event B is 0.700, and the probability of event B given event A is 0.200. Use Bayes Theorem to solve for the probability of even A given event B.

48. The probability of some event C is 0.060, the probability of some other event D is 0.950, and the probability of event D given event C is 0.500. Use Bayes Theorem to solve for the probability of even C given event D.

49. The probability of some event C is 0.060, the probability of some other event D is 0.950, and the probability of event D given event C is 0.500. Use Bayes Theorem to solve for the probability of even C given event D.

50. A patient goes to see his doctor complaining of blurred vision and thinks he has *dreaded yellow fickle-berry disease* (DYFBD). Based on the patient’s history, the doctor knows the probability of this patient having DYFBD is 0.100. The doctor orders a test, which comes back positive for DYFBD with a probability of 0.300. The doctor also knows that if the patient does have DYFBD, the probability of the test being positive is 0.900. The test comes back positive for DYFBD. Using Bayes Theorem, what is the probability that the patient has DYFBD, given the positive test result?

51. A patient goes to see his doctor complaining of blurred vision and thinks he has *dreaded yellow fickle-berry disease* (DYFBD). Based on the patient's history, the doctor knows the probability of this patient having DYFBD is 0.200. The doctor orders a test, which comes back positive for DYFBD with a probability of 0.300. The doctor also knows that if the patient does have DYFBD, the probability of the test being positive is 0.900. The test comes back positive for DYFBD. Using Bayes Theorem, what is the probability that the patient has DYFBD, given the positive test result?

52. Compute the following permutations and combinations:

a. ${}_5P_2$ b. ${}_5C_2$ c. ${}_6P_2$ d. ${}_4C_4$ e. ${}_4P_3$

f. ${}_6C_3$ g. ${}_4P_4$ h. ${}_3C_2$ i. ${}_3P_2$

53. You conduct a study that involves showing participants six clips. You are concerned that the order in which you present the movie clips could affect the outcome, so you decide to present them to each person in a different order. In how many ordered sequences can the movie clips be presented?

54. From #53, you decide that you can only show three movie clips to each subject. How many different ordered sequences of three movie clips of six movie clips can be presented?

55. You are interested in the effects of five independent variables on a dependent variable, but, you can study only two independent variables. How many combinations of two variables could you study?

56. From #55, how many combinations of three independent variables are possible?

57. A company has developed ten hot sauces, but only wants to market the three best-tasting sauces. To identify the best-tasting sauces, the company has people taste the hot sauces and give ratings. But, the company only wants to have each person taste four of the ten sauces. How many combinations of four hot sauces can be administered across people?

58. A team of three people from a widget company to assess the widget needs of a client is formed from a group of 3 managers, 10 analysts, and 15 technicians.

a. What is the probability that the team is composed of only analysts?
b. What is the probability that the team is composed of only technicians?
c. What is the probability that the team is composed of only managers?
d. What is the probability that the team is composed of two managers and the third person is either an analyst or a technician?
e. What is the probability that the team is composed of two analysts and one technician?

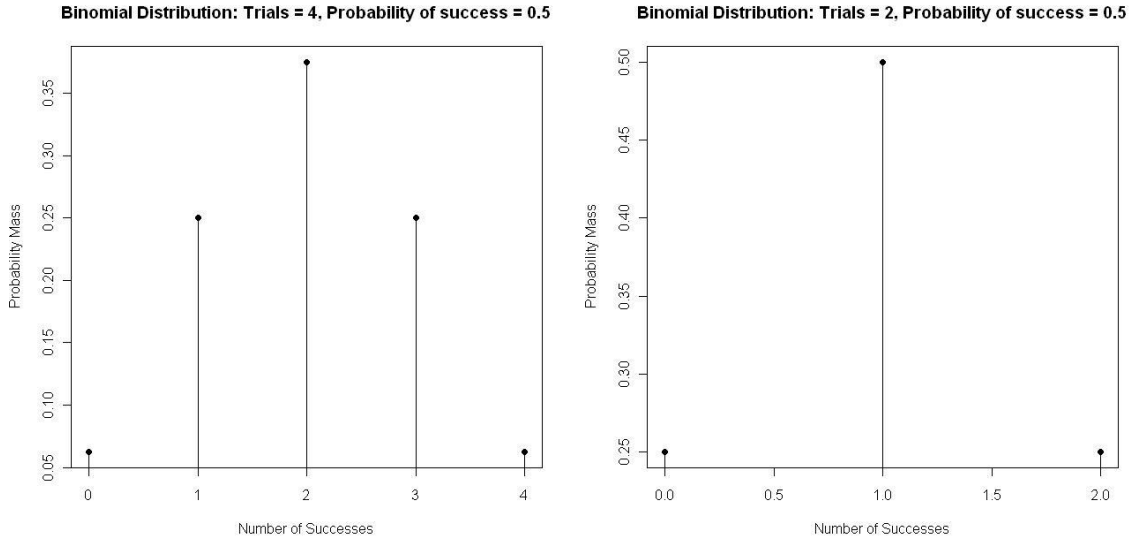
Chapter 9: Binomial Probability and Estimation

The comic to the right illustrates the point of this chapter which is determining the probability of a particular outcome of a random experiment when only two events are possible on each trial in a series. You may be thinking, '*wait, with only two events possible, assuming each event is equally probably, isn't the probability of either event 0.5?'* The answer is basically yes. When you have two possible outcomes on any trial (e.g., heads and tails) and each outcome has an equal probability, then the probability of observing either outcome on any one trial is 0.5. But, this chapter addresses the probability of a number of events over a number of trials, where the resulting probability is not 0.5.

9.1 Binomial Probability Expression

Remember, a simple random experiment is any procedure where the outcome is unknown in advance. A **binomial experiment** is a type of simple random experiment where only two mutually exclusive outcomes are possible on any trial and those two outcomes are a **success** and **failure**. Such trials where only one of two mutually exclusive outcomes is possible are **Bernoulli trials**. For example, flipping a coin is a Bernoulli trial, because only heads and tails are possible. Heads could be defined as a "success" and tails could be defined as a "failure." Rolling a six-sided die could be a Bernoulli trial if a roll of six was defined as "success" and a roll of any other number was a "failure." Finally, a person with cancer who is taking a new experimental type of chemotherapy is a Bernoulli trial, where the patient being cured is a "success" and the patient not being cured is a "failure."

The **binomial probability** is the probability of observing a certain number of successes (r) over a certain number of independent Bernoulli trials (n). The **binomial probability distribution** is the theoretical probability distribution of all numbers of possible successes over a certain number of Bernoulli trials. Below are two examples of binomial probability distributions (for n = 2 trials and n = 4 trials) where the probability of success on any on trial is p = .50:



We'll work more with the binomial distribution later. Just for now note that it is a theoretical probability distribution that displays the probability of each number of successful outcomes, from 0 – n, in a binomial experiment, assuming a certain probability of success on each trial.

Consider the following example: You have an unweighted ("fair") six-sided die that when rolled should result in one dot, two dots, three dots, four dots, five dots, or six dots on the face side with equal probability. That is, the probability of observing a certain number of dots on any one roll of the die is p = 1/6 = 0.167. You roll the die and you try to guess how many dots will be showing. Because the die is unweighted the probability you will be correct ("success") on any one roll is p(Correct) = 0.167, and the probability you will be incorrect on any one roll of the die is p(Incorrect) = 1 - 0.167 = 0.833.

You roll the die ten times, guessing the number of dots before each roll. What is the probability you correctly predict the number of dots showing on zero of the ten rolls? How about the probability of being correct on one out of the ten rolls? How about the probability of being correct on two out of ten rolls? Many believe the probability of being correct a certain number times out of some larger number of trials is equal to the probability of being correct on any one trial raised to r^{th} power. That is, is p(correct on r trials) = p^r This is not correct for two reasons:

First, this can be used only to obtain the probability of a certain number of outcomes in a row. That is, if we were interested in the probability of rolling three dots on a die five times in a row, the probability of rolling five threes in a row is 0.167^5 = (.167)(.167)(.167)(.167)(.167) = 0.000129. But we want to know the probability of correctly guessing the face of the die on a certain number of trials out of ten; and those correctly-guessed trials could be sequential or spaced out. That is, there are many different combinations of a certain number of successes out of a larger number of events.

Second, this does not take into account the total trials (rolls of die). That is, we want to know the probability of a certain number of successful guesses of the die roll out of a certain number of trials. But, p^r assumes that the number of successes we are interested in is equal to the total number of trials. Of course, we might want to know the probability of five successes out of 10 trials; p^r cannot help up calculate that probability.

Raising the probability of being correct on any trial to a power will not suffice. Instead, we use the **binomial probability expression**, which is used to calculate the probability of *r successful outcomes over n trials, given the probability of success on any one trial is p*:

$$p(r \mid n, p) = \left[\frac{n!}{(n-r)!\,r!} \right] p^r q^{n-r}$$

In the binomial probability expression, r is the number of successful outcomes over the n total Bernoulli trials. The p is the probability of being successful on any one trial; and q is the probability of failure (of being unsuccessful) on any one trial, where $q = 1 - p$. Notice that the term inside of the brackets is the formula for determining the number of combinations of a subset of events out of a total number of events.

From the example above: let's say we want to know the probability of correctly predicting five rolls out of ten rolls of a die; hence r = 5 and n = 10. The probability of successfully predicting a roll of the die on any one trial is p = 1/6 = 0.167. This is so, because there are six sides of the die and the probability that you correctly guess the side that is showing after the roll, by chance, is one in six. Thus, the probability of failing to predict the roll on any one trial is $q = 1 - 0.167 = 0.833$. Plugging these values into the binomial probability expression from above, we have:

$$p(5 \mid 10, 0.167) = \left[\frac{10!}{(10-5)!\,5!} \right] (0.167)^5 (0.833)^{10-5}$$

$$p(5 \mid 10, 0.167) = \left[\frac{3628800}{(120)120} \right] (0.000129)(0.401074)$$

$$p(5 \mid 10, 0.167) = [252](0.000129)(0.401074)$$

$$p(5 \mid 10, 0.167) = 0.013$$

The probability of correctly predicting five of ten rolls of a die is 0.013, or the percent chance of accomplishing this feat is about 1.3%. Let's say we now want to know the probability of being correct on *r* = 6 out of the *n* = 10 trials. We have:

$$p(6 \mid 10, 0.167) = \left[\frac{10!}{(10-6)!6!}\right](0.167)^6(0.833)^{10-6}$$

$$p(6 \mid 10, 0.167) = \left[\frac{3628800}{(24)720}\right](0.000022)(0.481482)$$

$$p(6 \mid 10, 0.167) = [210](0.000022)(0.481482)$$

$$p(6 \mid 10, 0.167) = 0.002$$

Notice, as the number of successes increases, the binomial probability decreases. This is because as the number of successful outcomes out of a finite number of trials increases, the likelihood of observing that number of successes necessarily decreases. The table below shows the probability of being correct for each of the r possible successes out of n trials. That is, given ten rolls of a die, you can successfully predict the die roll on 0, 1, 2, 3,...,9, or 10 of those rolls:

Number of Correct Predictions (r)	**Probability**	**Cumulative Probability**	**Probability of being correct *r* times or more**
10	.00000002	1.00000000	.00000002
9	.0000008	.99999998	.00000086
8	.00002	.99999914	.0000197
7	.0002	.99998	.00027
6	.0022	.9997	.0025
5	.0131	.9975	.0156
4	.0546	.9844	.0702
3	.1555	.9298	.2257
2	.2909	.7743	.5166
1	.3225	.4834	.8391
0	.1609	.1609	1.0000

In the table above, the values in the *Probability* column are the probabilities associated with observing exactly that number of successful outcomes out of ten. The values in the *Cumulative Probability* column are the probabilities associated with observing that many successful outcomes *or less*. For example, the probability of correctly predicting r = 4 rolls of the die or less is 0.9844. The values in the last column (*Probability of being correct r times or more*) are the probabilities associated observing that many successful outcomes or more. For example, the probability of correctly predicting r = 8 rolls of the die or more is 0.0000197.

The binomial expression can be used in hypothesis evaluation. **Hypothesis testing** in statistics is all about determining whether the probability of observing some outcome. Generally, if the probability of observing something is very low then it is unlikely a random event and something must have caused that outcome.

For example, in the table above you can see that the probability of correctly predicting eight rolls of the die out of 10 rolls is *p* = 0.00002, which is a very low probability. Intuitively, one would assume it is highly unlikely that a person would correctly predict the roll of a die that many times if the person was just guessing. Because the probability of such an occurrence is so low, if someone does accomplish this feat, one might conclude that the person has *precognition* regarding the outcome of a roll of a die and is able to predict how many dots will be showing on the roll. (Check out Daryl Bem's work on precognition)[1]

[1] Some websites on precognition, which does not exist: http://dbem.ws/
http://psychsciencenotes.blogspot.com/2010/11/brief-note-daryl-bem-and-precognition.html

As another example of how the binomial expression can be used in in hypothesis testing, let's say we have a coin and we are trying to determine whether the coin is weighted (biased) or unweighted (unbiased). Our initial hypothesis should be that the coin is unbiased and will equally often show heads and tails. The reason for this initial hypothesis is that we have no evidence to suggest that the coin is biased, and because the coin can only show one of two sides when flipped (heads of trials), the probability of showing heads should be about ½ and the probability of showing tails should be about ½.

To determine whether the coin is based, we decide to perform a random experiment by flipping the coin ten times We consider a flip of heads to be a successful outcome (we could also do this for tails), so we count the number of heads that show up over the ten flips. Under our initial hypothesis, if we assume the coin is unbiased, the probability of heads on any one flip of the coin is p(heads) = 0.5; thus, if we flip the coin n = 10 times we should expect, by chance, 10 x 0.5 = 5 heads. That is, if the coin is unbiased and the probability of heads on any one flip is the same as the probability of a tails on any one flip, then we should expect an equal number of heads and tails over the after 10 flips of the coin.

Let's say we end up tossing six heads. Does this mean that the coin is biased toward showing heads? Maybe, but how likely is this outcome of six heads? This is where the binomial probability expression comes in. If we calculate the probability of observing r = 6 heads over n = 10 flips of the coin (p = 0.5 and q = 0.5), we find that:

$$p(6 \mid 10, 0.5) = \left[\frac{10!}{(10-6)!6!}\right](0.5)^5(0.5)^{10-5}$$

$$p(6 \mid 10, 0.5) = \left[\frac{3628800}{(24)720}\right](0.03125)(0.03125)$$

$$p(6 \mid 10, 0.5) = [210](0.03125)(0.03125)$$

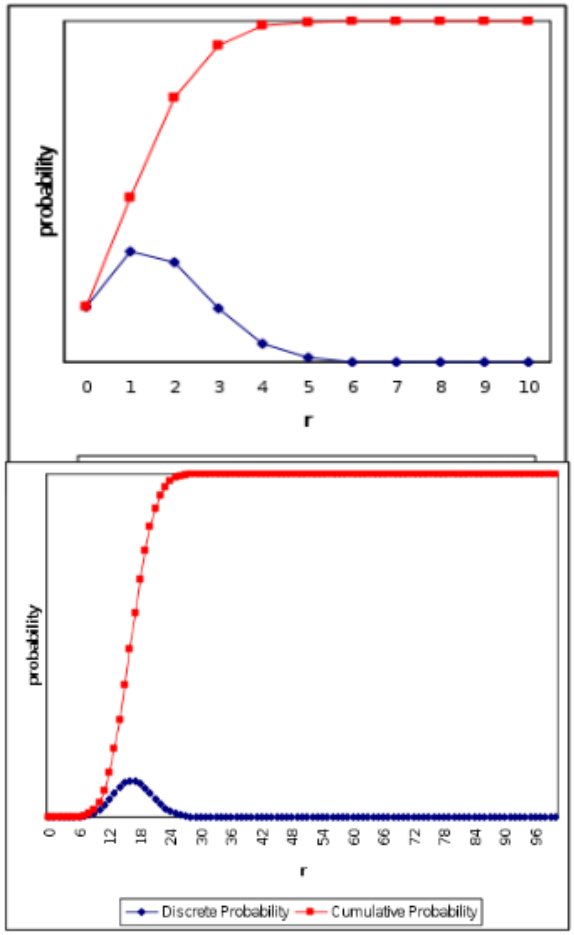
$$p(6 \mid 10, 0.5) = 0.205078$$

There is roughly a 20% chance of tossing six heads over 10 flips of the coin, if the coin is unbiased. Thus, there is pretty good chance of observing this outcome, even if the coin is unbiased. As such, we decide to stick with our original hypothesis that the coin is unbiased, because there is not substantial evidence to suggest that the coin is biased.

Generally, an outcome to be considered *statistically significant* enough for us to change our original hypothesis, the probability of the outcome must be equal to 0.05 or less. That is, if the binomial probability comes out to be 0.05 or less, then the probability of that outcome is so low that the initial hypothesis is not likely correct, so we should change it. But, this is not the case here.

What if we had thrown nine heads? Would we reject our initial hypothesis that the coin is unbiased and change our hypothesis to the coin is biased? In this case, the probability of r = 9 heads over n = 10 flips if the probability of heads on any one trial is .5 is p = .009766, which is quite low. In this case, because the probability of this outcome is so unlikely if the coin was unbiased, we would reject that hypothesis and conclude the coin is actually biased.

9.2 Binomial Probability Distribution



The graph to the right plots the binomial probability of observing *r* successful outcomes out of *n* trials, and the cumulative binomial probability *r* successful outcomes or less out of *n* trials. The plots are based on the die-rolling example in the previous section. The abscissa is the number of successful predictions of a die roll from *r* values of zero to 10 and the ordinate is the probability of each value of *r*. This is an example of the binomial probability distribution (blue line) and a cumulative binomial probability distribution (red line) for *n* = 10, *r* = 0 to 10, and *p*(success) = 0.167.

Importantly, as you increase the size of *n* the binomial probability distribution approximates a normal distribution. You can see this in the blue line in the graph to the right, which presents the binomial probability distribution for all values of *r* from 0 to n = 100, and *p*(success) = 0.167. Notice the peak of the binomial probability distribution (blue line) is shifted to the left of the center on the abscissa, even though the shape of the distribution is "normal." On close inspection you might notice that the peak is at the value that you would get by multiplying the probability of successfully predicting the roll of a die on any one trial [p(success) = 0.167] by the total number of trials (n = 100). That is, the peak is near the value 0.167 x 100 = 16.7. This is because with a sufficiently large *n* the binomial probability distribution will approximate a normal distribution with a mean that is equal to p x n; a topic I will turn to in the next section.

9.3 Binomial Approximation to the Normal Distribution

With a sufficiently large *n* (Bernoulli trials) the binomial probability distribution will approximate a normal distribution. This **binomial approximation of the normal distribution** has a mean equal to:

$$\mu = np$$

This is the expected number of successful outcomes expected by chance. In the expression, *n* is the number of Bernoulli trials and *p* is the probability of a successful outcome on any one trial. The variance and the standard deviation of the binomial distribution are:

$$\sigma^2 = npq \qquad\qquad \sigma = \sqrt{npq}$$

In each expression, *n* is the total number of trials, *p* is the probability of success on any one trial, and *q* is the probability of failure on any one trial. For example, say that we roll an unweighted die n = 1,000 times. We would expect the mean, variance and standard deviation to be:

$$\mu = (1000)(0.167) = 167$$

$$\sigma^2 = (1000)(0.167)(0.833) = 139.111$$
$$\sigma = \sqrt{(1000)(0.167)(0.833)} = \sqrt{139.111} = 11.794$$

Because we know the mean (μ) and standard deviation (σ) of this binomial distribution, we can calculate a *z*-score for any conceivable value of *r* from 0 to *n* and then use that *z*-score to determine different probabilities associated with *r*. For example, using the mean and standard deviation from above, say we want to know the probability of successfully predicting 200 rolls *or more* rolls of the die. We first calculate the z-Score for *r* = 200:

$$z = \frac{X - \mu}{\sigma} = \frac{200 - 167}{11.794} = \frac{33}{11.794} = 2.798$$

Next, we look up this z-score in the standard normal tables (Table 1 in the statistics packet and Appendix A). Once we find the z-score, we use the values in Column 3 to determine the probability of being correct on 200 rolls of the die *or more*, which you will find is *p* = .0026. This probability can be used in hypothesis testing.

For example, say we were trying to determine whether a person has precognitive abilities (extrasensory perception, ESP), and we test their precognitive abilities by having them guess the rolls of a die. Our initial hypothesis is the individual does not have precognition, because there is no reason to believe the individual does have precognitive abilities. To determine whether the individual has precognitive abilities we roll the unweighted die 1000 times and have the individual predict each roll. The individual is correct on 190 of the rolls. As a criterion, we will assume that if the probability of being correct this many times or more is less than .05, we will conclude that the individual has precognitive abilities. That is, if the probability of correctly predicting 190 of 1000 rolls of the die is .05 or less, we will consider this number to be statistically significant and we will change our original hypothesis. First, we calculate the z-score that is associated with r = 190 successful predictions:

$$z = \frac{X - \mu}{\sigma} = \frac{190 - 167}{11.794} = \frac{23}{11.794} = 1.950$$

Like before, look up this z-score in the standard normal tables. Once we find the z-Score we use Column 3 to determine the probability of being correct on 190 rolls of the die *or more*, which we find is *p* = 0.0516, which is just slightly greater than our criterion probability of .05. Thus, because 0.0516 > 0.05, we will conclude that correctly predicting 190 rolls of a die out of 1000 is not significantly different from predicting the expected (mean) number of rolls at 167. Therefore, we will not change our original hypothesis and we conclude that there is insufficient evidence to indicate this individual has precognitive abilities.

For another example, say we want to know the cumulative probability of successfully predicting between zero and 175 rolls of the die, that is, of being correct on *at most* 175 rolls. First, calculate the z-Score for *r* = 175:

$$z = \frac{X - \mu}{\sigma} = \frac{175 - 167}{11.794} = \frac{8}{11.794} = 0.678$$

Look up this z-Score in the standard normal tables and use the value in Column 5 to determine the probability of being correct between μ = 167 and r = 175 rolls of the die, which you will find is *p* = 0.2517. (This is the probability between the mean and the z-score.) Finally, add that probability to 0.5, which will give you the cumulative probability of correctly predicting *at most r* = 175 rolls of the die: p = .2517 + .5000 = .7517.

CH 9 Homework Questions

1. A student is taking an eight-question true-false test. If the student randomly guesses on each question, what is the probability that this student:

a. gets eight correct
b. gets six correct
c. gets four correct
d. gets one correct

2. You have an unbiased six-sided die that you roll ten times, and record the number of dots on the face side. You consider a roll of thee dots to be a 'success' and any other number of dots to be a 'failure'. What is the probability of rolling three dots:

a. eight times
b. two times
c. four times
d. five times

3. Using the same die and information in Exercise 2, say that you roll the die 10,000 times and count the number of 3's (successes). What is the expected (mean) number of 3's that you will roll after 10,000 rolls of the die? What is the variance and standard deviation of the binomial distribution based on n = 10,000 rolls?

4. From Exercise 3, say that you roll a three 1750 times. What is the z-Score of this number of 3's? What is the probability of rolling this many three's or more?

5. Calculate the mean, variance and the standard deviation for a binomial distribution with n = 150 trials and the probability of success on any trial is 0.600.

6. Using the information in Exercise 5, determine the probability of:

a. 100 successes or more:
b. 110 successes or more:
c. 84 successes or fewer:

7. Calculate the mean, variance and standard deviation of a binomial distribution with n = 200 and p = 0.800.

8. Using the information in Exercise 7, determine the probability of:

a. 175 successes or more:
b. 140 successes or fewer:

9. A student takes a multiple-choice quiz of ten items. Each item has four response options. The student gets five items correct. Using an alpha level of $\alpha = .05$ as a criterion, did the student perform at an 'above chance' level?

10. From #9, what if the student had gotten six items correct. Using an alpha level of $\alpha = .05$ as a criterion, did the student perform at an 'above chance' level?

11. Dr. Claws'n'paws is a biology professor who teaches a lab in biologic evolution. The lab includes 12 students: 8 female students and four male students. One would assume that the probability that male and female students enroll in this lab is p = .50 Using this information calculate the binomial probability of having this many female students enroll in the lab. When you are calculating your answer, be sure to do any calculations out to as many decimal places as necessary to keep your answer non-zero. Is the probability of having this many females in the lab likely, or unlikely, due to chance?

12. The final exam in your Cyberporn and Society class (look it up, it exists!) contains 25 questions, each with four answer to choose from. Assume that you did not study for this exam and you end up selecting one of the four answer choices to each question at random. You need to get 60% of the 25 questions correct in order to pass the exam. Use this information to answer the following:

a. By chance, what is the probability of selecting the correct answer on any one question?
b. What is the minimum number of questions that you need to answer correctly to pass the exam?

13. The possible letter grades that you can obtain in any class are (from high to low): A, A-, B+, B, B-, C+, C, C-, D+, D, and F. Both psychology and neuroscience students must take PSYC 330 (Research Methods), which requires a grade of C or higher in PSYC 210 (Statistics). Hence, a grade of C or higher should be considered a "success" and a grade of C- and lower should be considered a "failure" for being admitted into PSYC 330. Let's say that I go a little crazy one year and start awarding course grades to my PSYC 210 students at random (i.e., I pull a student's grade out of a hollow glass head).

a. For any one student, what is the probability that I will select a grade that is high enough to get into PSYC 330?
b. Say that I teach 150 students in PSYC 210 in an academic year. How many students should I expect will get a grade that is acceptable for PSYC 330?
c. Calculate the variance of the binomial probability distribution.
d. Calculate the standard deviation of the binomial probability distribution.
e. What is the probability of this many students receiving a grade of C or higher?

14. I have a Canadian Toonie (a $2 CD coin). Honestly, I have flipped one of these and it has landed on its edge. Assume that the probability of showing a heads is 0.49, the probability of a tails is also 0.49; hence, the probability of landing on its edge is 0.02. You consider a flip of the Toonie landing on its edge to be a "success."

a. If I flip this Toonie 10,000 times, what is the expected number of times the Toonie should land on its edge?
b. Calculate the binomial probability variance.
c. Calculate binomial probability standard deviation.
d. Say the Toonie lands on its edge 180 times. What is the probability that the Toonie lands on its edge 180 times out of these 10000 flips?

15. Twenty percent of individuals who seek psychotherapy will recover from their symptoms irrespective of whether they receive treatment (a phenomenon called spontaneous recovery). A researcher finds that a particular type of psychotherapy is successful with 30 out of 100 clients. Using an alpha level of .05 as a criterion, what should she conclude about the effectiveness of this psychotherapeutic approach?

Chapter 10: Introduction to Hypothesis Testing

10.1 Inferential Statistics and Hypothesis Testing

Earlier chapters dealt mainly with descriptive statistics, and the remaining chapters deal mainly with inferential statistics and hypothesis testing. Inferential statistics are analyses conducted on sample data, the outcomes of which can be used to make inferences about what would likely be found in a population. The picture of the scale above is to emphasize that hypothesis testing is all about weighing evidence, and whether the weight of the evidence is sufficient enough for an outcome to be *statistically significant*.

For example, say we provide the nutritional supplement ginko-baloba to a sample of college freshmen and give a placebo to another sample of freshmen. We find the freshmen who took ginko-baloba have a higher mean GPA at the end of the freshman year compared to students taking the placebo; thus, there appears to be a relationship between taking ginko-baloba versus a placebo and GPA. To assess this relationship we use inferential statistics to determine whether difference in GPA between students taking ginko-baloba and students taking the placebo is **statistically significant**. Simply put, two things may be numerically different, but it does not mean that this difference is statistically different.

Inferential testing boil down to a **test-statistic**, which is a ratio of the size of a relationship or **effect** to some measurement of sampling **error**. Thus, any test statistic can be summarized as:

$$\text{test statistic} = \frac{\text{effect}}{\text{error}}$$

We generally want the ratio of effect to error to be large, which would indicate the observed effect or relationship is greater than random sampling error. The question addressed in this and following chapters is how large the ratio needs to be for a relationship to be statistically significant. First, remember that in any study one must explicitly state the predictions of their null and alternate hypotheses. Recall, the **null hypothesis** (H_0) predicts the expected relationship will <u>not</u> be observed and the **alternate hypothesis** (H_1) predicts the expected relationship <u>will</u> be observed. From the ginko-baloba example above, one might predict that taking ginko-baloba, which some believe affects learning and memory, will affect student GPAs.

In this case the null hypothesis would predict that students taking ginko-baloba will have GPAs that are different than the GPAs of students taking a placebo; and the alternate hypothesis would predict that students who are taking ginko-baloba will have GPAs that do differ from those of students who are taking a placebo. When stating hypotheses, symbols are used to state the null and alternate hypotheses, and they should be stated in terms of their populations. For example, in the ginko-baloba example above the null and alternate hypotheses would be written symbolically as (be sure to use subscripts):

$$H_0: \mu_{Ginko} = \mu_{Placebo}$$
$$H_1: \mu_{Ginko} \neq \mu_{Placebo}$$

Both hypotheses symbolically state the predictions described in the preceding paragraphs, with the null hypotheses stating that taking ginko-baloba will not lead to a change in GPA relative to taking a placebo. In contrast, the alternative hypothesis is stating there will be a difference in GPA between ginko-baloba and taking a placebo. The point is when stating hypotheses, you should write them symbolically rather than verbally, because meanings of symbolic statements are universal. You may wonder why population symbols are used rather than sample symbols. Remember, inferential statistics use sample data to make inferences about what should be found in the population from which the sample came; thus, hypotheses should reflect the inferences, which is population symbols are used.

10.2 Null Hypothesis Significance Testing (NHST)

The logic to **null hypothesis significance testing** (**NHST**) may seem a little strange, convoluted, and flawed at first. Indeed, there are some who are critical of hypothesis testing and believe it should be banned. Null hypothesis significance testing is a tool for making inferences from sample statistics about unknown parameters; and .when used properly, NHST is very effective. Unfortunately, some people do not take into account many of the parameters of NHST and decide a relationship exists when the relationship is weak.

In NHST, the null hypothesis, not the alternate hypothesis, is assessed. This seems strange, because the alternate hypothesis is usually the one a researcher wants to confirm; hence, it would seem logical to test the alternate hypothesis. But if this was so, then a researcher would be seeking evidence only to confirm predictions and may become biased in evidence-gathering. In scientific reasoning you want to disconfirm objections to your prediction; thus, one should seek to disconfirm their null hypothesis. If you can disconfirm the null, you can use the alternate in its place (I know this already sounds weird...bear with me).

In NHST, you start by *assuming* the null hypothesis is true, that is, without any evidence to the contrary you assume there should be no relationship between the variables under study. This is synonymous with in the previous chapter where we started out by assuming that a coin was unbiased, because we had no reason to believe it was biased.

You then collect data and apply the appropriate inferential statistics to determine the probability of observing the outcome, while still assuming the null hypothesis is true. The inferential statistic tell us p(Outcome | H_0 = True), that is, the probability of observing our result if the null hypothesis was true. Critically, p(Outcome | H_0 = True) does *not* tell you the probability the null is true; it is the probability of observing the result assuming the null is true. If the probability of observing a particular outcome given the null is true is low, the null hypothesis is rejected. If p(Outcome | H_0 = True) is low, this means obtaining the outcome is unlikely if the null hypothesis was actually true. Stated differently, *if the probability of observing data is very low (assuming the null is true), then the null is unlikely correct and we reject the null hypothesis*. Once the null hypothesis is rejected, we accept the alternate hypothesis.

But at what point is p(Outcome | H_0 = True) low enough that the null hypothesis can be rejected? How low does the probability of an outcome have to be before we can say that the null hypothesis is unlikely to be true? The accepted level of **statistical significance** is $p = .05$ *or less*, which is called your **alpha-level** (**α**). Assuming the null hypothesis is true, if the probability of observing an outcome is $p = .05$ or less then it is unlikely this result would be obtained if the null was true, so the null is rejected. Researchers use different

alpha levels for different reasons, but the generally accepted level of statistical significance is α = .05 or less. Important: The alpha level is not the probability that the null hypothesis is true or false; it is the probability of observing an particular outcome *given* the null hypothesis is true.

10.3 Example with Binomial Expression

Recall, the binomial probability expression is used to calculate the probability of observing some specified number of successful outcomes (r) out of some number of trials (n). We have a six-sided die and want to determine whether it is weighted in such a way that four dots appear more often than the other sides. The null hypothesis in this example predicts that the number of fours rolled will be no different than other numbers of dots rolled, that is, the number of fours that are actually rolled will be equal to the number of rolls of fours that are expected by chance. If the die is rolled ten times and the *assumed* probability of rolling four dots on any one trial is $p = 0.167$ ($1/6^{th}$), we would expect (0.167)(10) = 1.67 rolls of four-dots. The alternate hypothesis predicts that the number of fours rolled will exceed the number that is expected by chance. The alternate hypothesis predicts that the number of rolls is greater than 1.67, because we are trying to see if the coin is biased to show the four-dot side, which would mean the four-dot side would snow more often. Thus, the hypotheses are:

H_0: $\mu_4 = 1.67$
H_1: $\mu_4 > 1.67$

We roll the die 10 times and record whether each roll results in a four (success, $p = 0.167$) or not (failure, $q = 0.833$). Say the four dots side shows up six times. In this case, the number of rolled fours and the number of rolled fours that is expected by chance (1.67) differ numerically. But, what we want to address is whether the six rolls of four-dots and the expected number of four-dots (1.67) differ statistically. Using the binomial expression, we calculate the binomial probability of rolling the four-dot side six times:

$$p(6 \mid 10, 0.167) = \left[\frac{10!}{(10-6)!6!}\right] 0.167^6 0.833^{10-6}$$

$$p(6 \mid 10, 0.167) = \left[\frac{3628800}{(24)720}\right] (0.000022)(0.481482)$$

$$p(6 \mid 10, 0.167) = [210](0.000022)(0.481482)$$

$$p(6 \mid 10, 0.167) = 0.002$$

The probability of rolling four-dots six times is 0.002. This is less than $p = .05$, which is the conventional alpha-level. In this case, we can say that the number of four-dots sides rolled is statistically greater than the number of four dot sides that were expected by chance. This is a statistically significant outcome and the null hypothesis can be rejected and the alternate hypothesis can be accepted. In layman's terms, we would conclude that that die is biased toward rolling four-dots.

10.4 The z-test

The **z-test** compares a sample mean to a population mean (μ) to determine whether the difference between means is statistically significant. To use the z-test, a sample must be drawn from a population with a known mean (μ) and standard deviation (σ), because the z-test calculates a z-score for the sample mean with respect to the population mean. The greater the difference between the sample mean and population mean, the larger the z-Score, and the less likely the sample mean is statistically equivalent to the population mean.

Assume we predict taking ginko-baloba will have an effect on freshman GPA. We randomly sample $n = 25$ freshmen taking ginko-baloba and at the end of their freshman year we ask for their GPA and we find this sample has a mean GPA of 3.20 (about B+). Let's say we also know that the mean GPA of all freshmen (including the 25 students in the sample) at this university is μ = 2.80 (about B-), with a standard deviation of σ = 1. Because we know μ and σ, we can use the z-test to determine whether the difference between the sample mean and μ is statistically significant.

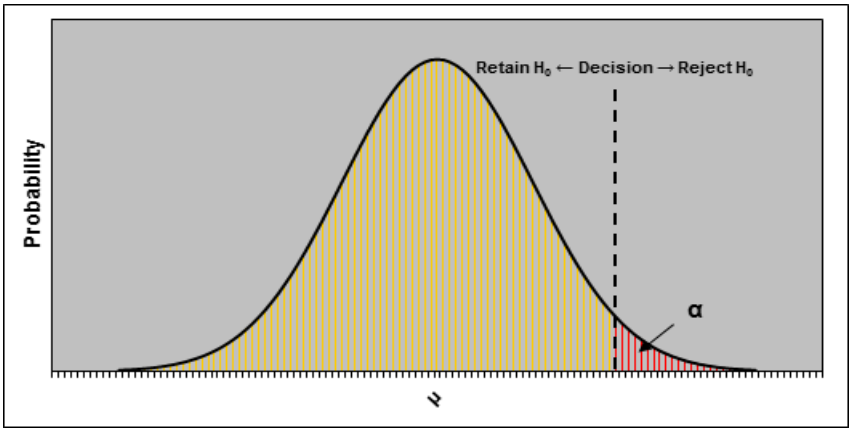
The first step in the z-test, which is the same for any inferential test, is the state your hypotheses. But first, I distinguish between **directional (one-tailed)** and **non-directional (two-tailed) hypotheses**, which affects how the null and alternate hypotheses are stated. Directionality is related to the alternate hypothesis: A directional alternate hypothesis predicts the sample mean will be, specifically, *greater than* or *less than* the population mean. Hence, the sample mean is predicted to lie in a specific direction relative to μ, that is, in one of the two tails of the distribution around a mean. In contrast, a non-directional alternate hypothesis predicts the sample mean will be *different* from the population mean. Hence, the sample mean could be greater-than or less-than μ, it doesn't matter; that is, the sample mean could fall into either one of the two tails of the distribution around a mean. For expository purposes, I go through this example using both types of hypotheses simultaneously.

From the ginko-baloba example above, a directional alternate hypothesis might predict that students who are taking ginko-baloba will have a mean GPA *greater* than the mean GPA of all freshmen (μ). A non-directional alternate hypothesis would state that the mean GPA of students taking ginko-baloba will be different from the mean of all freshmen. Symbolically, these hypotheses are:

Directional:	H_0: $\mu_{Ginko} = 2.80$	H_1: $\mu_{Ginko} > 2.80$
Non-directional:	H_0: $\mu_{Ginko} = 2.80$	H_1: $\mu_{Ginko} \neq 2.80$

Before continuing I must say something about the alpha level and directionality. Remember, the alpha-level is the probability associated with statistical significance, that is, the probability of observing an outcome assuming the null is true. In this example we'll use the conventional level of significance by choosing an alpha-level of α = .05. First, the alpha-level is a probability; it's the probability associated with statistical significance. Second, this probability is an area under a distribution, just like we covered in Chapter 6. Specifically, the alpha-level is an area under the tail end of a distribution.

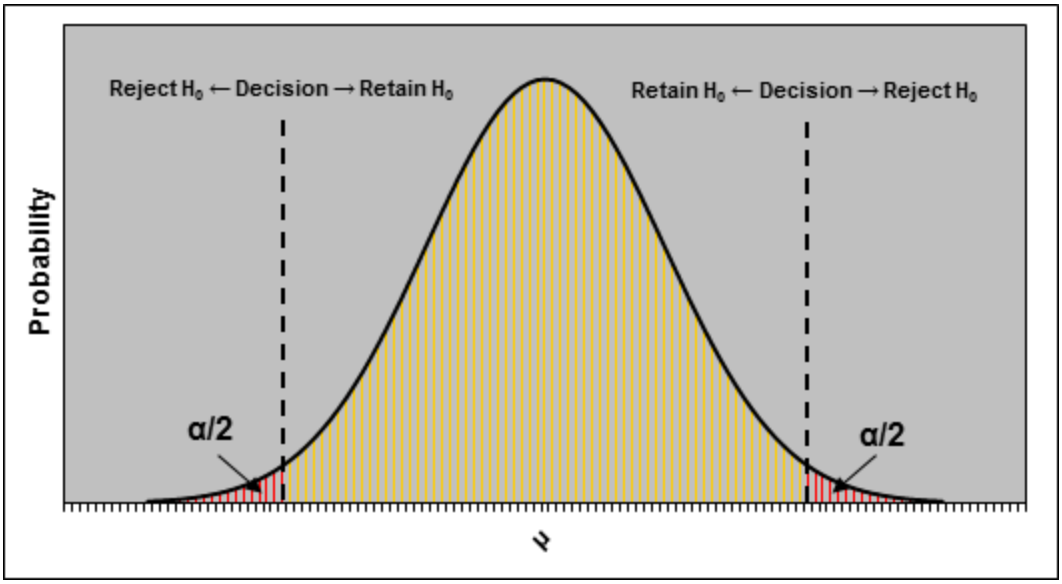
Here's where things get confusing. If you have a directional alternate hypothesis, which is also called a **one-tailed hypothesis test**, the alpha-level is an area in only one of the two tails of the distribution. But, if you have a non-directional alternate hypothesis, which is also called a **two-tailed hypothesis test**, the alpha-level is an area under both tails. For example, assume you predict a sample mean will fall in the upper tail of a distribution. In this case, the alpha-level will be associated with an area under the right tail of the normal distribution:



The probability of being in this area is equal to your alpha-level (α = .05 in this example) and this area is called the **critical region**. If the sample mean falls in this area we reject the null hypothesis and accept the alternate hypothesis and state there is a statistically significant difference between the sample mean and the population mean.

What about non-directional alternate hypotheses? In this case, the alpha-level remains the same value (α = .05 in our example), but this probability is split between the two tails, that is, the alpha-level is divided by two (α/2 = .05/2 = .025) and this probability is assigned to critical regions located in each tail (see below). The probability of being in one critical region is .025, but the probability of falling in either critical region

is .05. It is more difficult to be in a critical region for a non-directional test (the areas are smaller). This means a sample mean must be farther from the population mean for it to be significantly different from the population mean.



These concepts are important in determining statistical significance, because you must take directionality into account when assessing the outcome of an inferential test. Specifically, the probability associated with your test-statistic must be less than your chosen alpha level (.05) in order to claim there is a statistically significant difference between a sample mean and a population mean.

The z-test calculates the z-Score of a sample mean with respect to a population mean. The z-test is simply the difference between a sample mean and its population mean divided by the standard error of the sampling distribution of the mean (standard error of the mean):

$$z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}$$

From our ginko-baloba example, the sample mean was 3.20, the sample size was *n* = 25. The population mean was μ = 2.8, and the population standard deviation was σ = 1.0. The standard error of the mean in this example is:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{25}} = 0.2$$

Plugging the sample mean, the population's mean and standard error of the mean into the z-test, we obtain:

$$z = \frac{3.20 - 2.80}{0.2} = 2$$

This value (z = 2) is the **obtained value** (z_O) or **test statistic**, and tells us that the sample mean (3.20) is two standard errors above the population mean of 2.80. This test statistic value is what we use to determine whether the difference is statistically significant, and whether we should reject the null hypothesis.

Next, determine whether the outcome of the z-test (z_O) is statistically significant. To do this, look up the obtained value of z = 2 in the z-tables, a portion of which is reproduced, below with the relevant area with z = 2 highlighted in yellow. We need to determine the probability associated with z = 2. (We'll do this for both a directional and a non-directional alternate hypothesis.) For a directional hypothesis, look up the probability in column 3 (remember, column 3 provides the probability at or beyond a particular location on the x-axis). This is the ***p*-value** associated with your test-statistic (z = 2), and if this *p*-value (*p* = .0228) is less than your alpha-level (.05), the difference between the sample mean and the population mean is

statistically significant and the null hypothesis is rejected and the alternate hypothesis is accepted. Indeed, in this example, $p < \alpha$, so for a directional hypothesis the null would be rejected.

Column 1	Column 2	Column 3	Column 1	Column 2	Column 3
z	**p (0 < x ≤ +z) or p (0 > x ≥ -z)**	**p (x ≥ +z) or p (x ≤ -z)**	***z***	**p (0 < x ≤ +z) or p (0 > x ≥ -z)**	**p (x ≥ +z) or p (x ≤ -z)**
1.36	0.4131	0.0869	**1.96**	0.4750	0.0250
1.37	0.4147	0.0853	**1.97**	0.4756	0.0244
1.38	0.4162	0.0838	**1.98**	0.4761	0.0239
1.39	0.4177	0.0823	**1.99**	0.4767	0.0233
1.40	0.4192	0.0808	**2.00**	0.4772	0.0228
1.41	0.4207	0.0793	**2.01**	0.4778	0.0222
1.42	0.4222	0.0778	**2.02**	0.4783	0.0217
1.43	0.4236	0.0764	**2.03**	0.4788	0.0212
1.44	0.4251	0.0749	**2.04**	0.4793	0.0207
1.45	0.4265	0.0735	**2.05**	0.4798	0.0202

For a non-directional hypothesis, look up the probability in column 3 and then double that value (2 x .0228 = .0456). If this *p*-value ($p = .0456$) is less than half your alpha-level (.05), the difference between the sample mean and the population mean is statistically significant and the null hypothesis can be rejected and the alternate hypothesis can be accepted. (As a rule of thumb: *Whenever you have a statistically significant outcome, you reject the null hypothesis and accept the alternate hypothesis. Whenever you have a non-significant outcome you retain the null hypothesis only and do nothing with the alternate hypothesis*. It is a little more complicated than this, but we'll cover it in more depth in class.)

In this example, the *p*-values for both the directional (.0228) and non-directional (.0456) alternate hypotheses were less than the alpha level of .05. Thus, for either alternate hypothesis there was a significant difference between the sample mean and the population mean; hence, we can reject the null hypothesis and accept the alternate hypothesis. In layman's terms for both the directional and non-directional hypotheses, we conclude that students who are taking ginko-balboa have a significantly higher mean GPA at the end of their freshman year compared to the mean population GPA of all freshmen.

10.5 Reporting the z-test in the Literature

When writing up the results of a z-test in a manuscript for publication, several parameters need to be reported to convey just enough information to the reader: (1) the sample mean, (2) the population mean or the value that the sample mean is being compared to, (3) the obtained z-Value, (4) the standard error of the mean, and (5) the p-value. There are variants to exactly how the outcome of a z-test (or any inferential test) is up depending on the journal. But, here is how you would report just the results of the z-test: $z = 2.00$, $SEM = .20$, $p = .0228$ (two-tailed).[1] The *SEM* refers to the standard error of the mean. Below, I present an example of how the results of the non-directional z-test reported in Section 11.4 would be written up in a manuscript (this is based on APA 6th edition style):

Results

In the present study, $n = 25$ university freshmen who were taking ginko-baloba every day during their freshmen year were randomly selected from all students identified as taking ginko-baloba. Each student's GPA was recorded at the end of the freshman year, and this sample was found to have a GPA of $M = 3.20$. The mean GPA of all

[1] There are several ways of displaying the alpha level. This is one way of displaying the results assuming a non-directional alternate hypothesis.

university freshmen at the end of the same academic year was μ = 2.80 (σ = 1.0). A z-test comparing two means resulted in a statistically significant difference between the sample mean and population mean, z = 2.00, *SEM* = .20, *p* = .0228 (two-tailed). Thus, students taking ginko-baloba had a significantly larger GPA than all other freshmen.

10.6 Type I vs. Type II Errors

Is hypothesis testing perfect? Sadly, no and it's possible to make a mistake, because NHST is based on probabilities, not certainties. Remember, the chose α level is the probability of observing an outcome if the null is true; thus, α = .05 means there is a small chance you would observe the outcome with the null being true, but there is still a small chance you would observe this data even with the null being true. This is related to the "dance of the means" in earlier chapters: the value mean is variable, so you might obtain a statistically significant result simply due to random selection. In short, because hypothesis testing is based on probabilities of being correct or incorrect, there is always going to be some probability that your decision can be wrong.

In hypothesis testing, the null hypothesis is *rejected* or the null hypothesis is *retained*. Thus, there are two possible decisions that can be made about the null hypothesis, regardless of whether the null is actually true or false. Independent of your statistical decision, in reality (in the population) the null hypothesis may be true or may not be true. Remember, in hypothesis testing you use sample data to make inferences about what you would expect to find in a population. It is possible that you conclude a sample mean is statistically different from a population mean, but in reality there is no difference. Thus, if you able to test an entire population you may find that the null hypothesis is really true or is really false. Thus, there are four combinations of your two statistical decisions regarding the null hypothesis and two possible realities regarding the null hypothesis. These are illustrated in the table below:

	H_0 is Actually True in Population	**H_0 is Actually False in Population**
Decide to Reject H_0	Type I Error $p = \alpha$	Correct Decision $p = (1 - \beta)$ = Power
Decide to Retain H_0	Correct Decision $p = 1 - \alpha$	Type II Error $p = \beta$

From the table, you can see there are two situations leading to a correct decision and two situations leading to errors. A **Type I Error** is committed when the null hypothesis is rejected, but in reality the null hypothesis is true and should be retained. The probability of committing this error is equal to the selected α level. To decrease the chance of making a Type I Error, a smaller α can be selected (e.g., .01, .001, .0001). A **Type II Error** is committed when you fail to reject the null hypothesis when retain the null hypothesis, but in reality the null hypothesis is false and should be rejected. The probability of making a Type II Error is equal to something called **beta** (**β**). The probability of not making a Type II Error, that is, rejecting a false null hypothesis, is equal to 1 – β and this is the **power** of a statistical test. Generally, you want to have high power in your inferential test; that is, a high probability of correctly rejecting a false null hypothesis. Typically, Power should be .80 of more (up to 1.0).

10.7 Statistical Power

The power of a statistical test is the probability of rejecting a false null hypothesis. There are methods for calculating the value for β and power, which are covered in Appendix C and in later chapters, but for now we address how statistical power is influenced by several factors.

First, smaller α-levels decrease power. Specifically, as the value of α decreases the chance of making a Type I Error decreases, but so too does power. This is because smaller alpha-levels are associated with larger critical values, that is, you need a larger difference between means or a stronger effect to detect a significant result; thus, holding everything constant, a statistically significant difference between means is *less* likely to be detected as the critical value increases.

Second, holding everything constant, as sample size increases the power of a statistical test increases. Indeed, this is why many researchers prefer larger samples to smaller samples. As you increased the size of a sample, the sample mean becomes a better estimator of the population mean and the standard error of the mean decreases. Thus, as you increase the sample size, you decrease the standard error of the mean and increase the chance of detecting a significant result, because the test statistic should increase. If you increase the sample size you decrease the standard error of the mean, which is the denominator in the z-test; hence, you increase the obtained z-Value.

Third, the ***effect size***, that is, the mean difference between the population mean and the sample mean in the numerator also influences power. Generally, larger effect sizes are associated with more statistical power. Holding everything constant, as you increase the effect size you increase the obtained z-Value and you increase the chance of detecting a significant difference.

The fourth factor that influences power is error; that is, the standard deviation (σ). Larger standard deviations (more variance) lead to less power. This can be seen in the re-written z-test above. Holding everything else constant, as the standard deviation increases the denominator will increase and z_O will decrease; thus, you are less likely to detect a significant result. Hence, larger standard deviations (more error) results in less statistical power. Of course, it is unlikely that the population standard deviation will change that much, so there is little to worry about here. Indeed, the three other factors that affect power are more or less under your control.

10.8 Limits of z-test

There is a major limitation to the z-test, and that is it can be used only if both μ and σ are known, which is unlikely. In cases where the population mean is known, but the population standard deviation is not, if a sample is sufficiently large, it is acceptable to use the *estimated* population standard deviation in place of the true population standard deviation to calculate an estimated standard error of the mean. However, there is no 'real' criterion for how large a sample must be before it is "sufficiently" large. Another limitation is that the z-test only allows a comparison of a sample mean to a known population mean. In many cases you might have two sample means that you want to compare. With the z-test this is not technically possible (though I have seen several books try to get around this). Thus, the z-test is hampered, which is what statisticians developed something called the ***t-test***, which we turn to in the next several chapters.

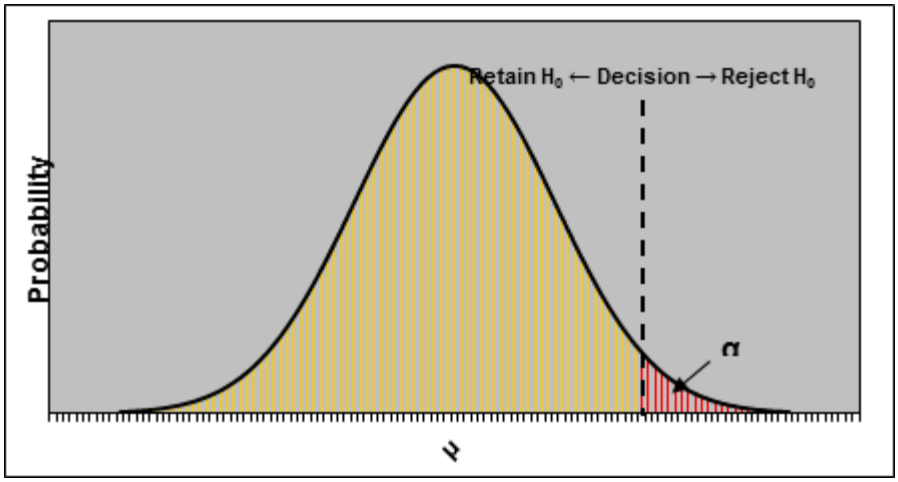
10.9 The "Classic" NHST Procedure

There is an alternative method for determining statistical significance to the method that was just described in the preceding sections. This is the "classic" method that is taught in nearly every undergraduate statistics book and assesses statistical significance by comparing the obtained test statistic to a pre-determined critical value. This section goes through this alternative approach, for historical value. The first step in this procedure is the same as before, that is, state the hypotheses:

Directional:	H_0: $\mu_{Ginko} = 2.80$	H_1: $\mu_{Ginko} > 2.80$
Non-directional:	H_0: $\mu_{Ginko} = 2.80$	H_1: $\mu_{Ginko} \neq 2.80$

Once the hypotheses are stated you determine what outcome is considered statistically significant, that is, what is the smallest test statistic that must be obtained for the outcome to be considered "statistically significant?" To do this you select an alpha-level and determine the necessary **critical values** for rejecting the null hypothesis. In this example we'll use the conventional alpha level of α = .05.The critical values are

the boundaries of the critical regions, that is, it is the point on the x-axis where the decision line falls for whether to reject the null hypothesis or retain the null hypothesis:



For directional hypotheses there is one critical value and one critical region and for non-directional hypotheses there are two critical values and two critical regions. Because the critical regions are areas associated with the selected alpha-level, critical values must be z-scores associated with the alpha-levels. The critical values are found in the standard normal table. Importantly, the **critical z-Value** (z_α) is determined by the alpha level (α) and whether you have a directional or a non-directional hypothesis.

From our example, to find the critical value for a non-directional hypothesis\m first divide the alpha-level in half: .05/2 = .025 and lookup this value in column 3 of Table 1, and use the z-score in column 1 as the critical value. The critical z-Value for a non-directional test with α = .05 is $z_\alpha = \pm 1.96$. The plus/minus is necessary, because the sample mean could be greater-than or less than the population mean.

To find the critical z-Value for a directional hypothesis: lookup the alpha-level (.05) in column 3 and use the z-score from column 1 as the critical value. From Table 1, $p = .05$ does not occur in column 3 and the two closest probabilities are .0505 and .0495. The rule of thumb is use the smaller of the probabilities. Thus the critical z-Value for a directional test with α = .05 is $z_\alpha = +1.65$. The plus/minus is not necessary for a directional test; however, because the sample mean was predicted to be greater than the population mean, you must place a plus sign in front of the critical value. If the sample mean was predicted to be less than the population mean, which means you are predicting a negative outcome in the z-test, then you must place a negative sign in front of the critical value. The next step is to conduct the inferential test. From the earlier sections, here are the standard error of the mean and the test statistic $\sigma_M = 0.2$, $z = 2$.

To determine whether this outcome is statistically significant, compare the test statistic (z = 2) to the critical value of $z_\alpha = \pm 1.96$. If the absolute value of the test statistic is greater than or equal to the absolute value of the critical value, the difference between the sample mean and population mean is statistically significant. Because z_α for the non-directional hypothesis in the ginko-baloba example is $z_\alpha = \pm 1.96$, the test statistic of z = 2 is greater than the critical value and difference between the sample mean and the population mean is statistically significant. Because z_α for a directional hypothesis was $z_\alpha = +1.65$, the test statistic of z = 2 is greater than the critical value; thus, the difference between the sample mean and population mean is statistically significant by the directional hypothesis.

In both cases, because the test statistic was greater than the critical value, the difference between the sample mean and the population mean is statistically significant. In this case, the null hypothesis can be rejected and the alternate hypothesis can be accepted. In future chapters, I do not cover this alternate method of assessing statistical significance.

CH 10 Homework Questions

1. Define each of the following:
    a. null hypothesis
    b. alternate hypothesis
    c. critical value
    d. rejection region
    e. test statistic
    f. alpha level

2. A high school administration is interested in whether their students save more or less than $100 per month toward college. Translate this question into a null hypothesis and an alternate hypothesis.

3. Why can we never accept the null hypothesis as being true based on statistical tests?

4. *Use the following information to answer the questions that follow*: H_0: μ = 5, H_1: μ ≠ 5, $\overline{X} = 8$, σ = 15, and n = 25.
    a. Calculate the standard error of the mean.
    b. Calculate the test-statistic for the sample mean.
    c. What is the p-value for the test-statistic?
    d. Using an alpha level of α = .05 test the viability of the null hypothesis.

5. *Use the following information to answer the questions that follow*: H_0: μ = 6, H_1: μ ≠ 6, $\overline{X} = 4.50$, σ = 3, and n = 36.
    a. Calculate the standard error of the mean.
    b. Calculate the test-statistic for the sample mean.
    c. What is the p-value for the test-statistic?
    d. Using an alpha level of α = .01 test the viability of the null hypothesis.

6. *Use the following information to answer the questions that follow*: H_0: μ = 6, H_1: μ < 6, $\overline{X} = 4.50$, σ = 5, and n = 49.
    a. Calculate the standard error of the mean.
    b. Calculate the test-statistic for the sample mean.
    c. What is the p-value for the test-statistic?
    d. Using an alpha level of α = .05 test the viability of the null hypothesis.

7. *Use the following to answer the questions below*: Scores on IQ tests are normally distributed with μ = 100 and σ = 15. I am interested in whether University of Scranton Psychology majors have a mean IQ greater than the general population. I randomly select n = 25 psychology majors and find they have a sample mean IQ of 104.
    a. Expressed in terms of μ, what are the null and alternate hypotheses?
    b. Calculate the standard error of the mean.
    c. What is the obtained z-Score for the sample mean of 104?
    d. What is the p-value associated with this test statistic?
    e. Using an alpha level of .05, what decisions should you make regarding the hypotheses? What should you conclude about the mean IQ of University of Scranton psychology majors compared to the general population?

8. *Use the following to answer the questions below*: You are a faculty member at East Buzzkill University and you believe students at your institution poorer reading comprehension skills than other college students. You test you hypothesis by comparing SAT-Verbal scores of students attending East Buzzkill University to the mean SAT-Verbal score of all students taking the SATs. Scores on the SAT-Verbal test are normally distributed with μ = 500 and σ = 75. You sample n = 100 East Buzzkill University students and ask for their SAT-Verbal scores. The sample has a mean of mean of 480.
    a. Expressed in terms of μ, what are the null and alternate hypotheses?
    b. Calculate the standard error of the mean.
    c. What is the obtained z-Score for the sample mean of 480?

d. What is the p-value associated with this test statistic?
e. Using an alpha level of .05, what decisions should you make regarding the hypotheses? What should you conclude about the SAT verbal scores and reading comprehension abilities of East Buzzkill University students?

9. *Use the following to answer the questions below*: The Diagnostic Algebra Test (DAT) is a test given to all incoming freshmen to test their general knowledge of algebra. Scores on the DAT are normally distributed with $\mu = 13$ and $\sigma = 3$. You believe students entering as freshmen Neuroscience majors score better than average on the DAT, so you randomly sample the DAT scores from $n = 10$ freshmen Neuroscience major. You fund this sample mas a mean of 13.8.

a. Expressed in terms of μ, what are the null and alternate hypotheses?
b. Calculate the standard error of the mean.
c. What is the obtained z-Score for the sample mean of 13.8?
d. What is the p-value associated with this test statistic?
f. Using an alpha level of .01, what decisions should you make regarding the hypotheses? What should you conclude about the mean DAT score of incoming Neuroscience majors?
e. Say you increase to $n = 100$. Recalculate the standard error of the mean.
f. Based on this standard error, recalculate the test-statistic.
g. What is the p-value associated with this test statistic?
h. Based on this new information and using the same alpha level, would your conclusion made in 'f' change?

10. Define each of the following.

a. Type I error
b. Type II error
c. alpha-level
d. beta
e. power

11. What is the relationship between alpha and the probability of a Type I error. What is the reason for this relationship?

12. What is the relationship between power and the probability of a Type II error? What is the reason for this relationship?

13. What is the relationship between alpha and power? What is the reason for this relationship?

14. What effects does sample size have on the power of a statistical test?

15. Under what circumstance should a directional rather than a non-directional test be used? Why? Under what circumstance should a non-directional rather than a directional test be used? Why?

16. What effect does sample size have on the power of a statistical test? Why?

Chapter 11: One-Sample t-test

11.1 Where and Why the one Sample t-test Exists

Chapter 10 covered the basics of hypothesis testing by introducing the z-test, and the end of that chapter addressed a limitation of the z-test: it can be used only when μ and σ are known. In cases where μ and σ are unknown, the **one-sample t-test** is used.

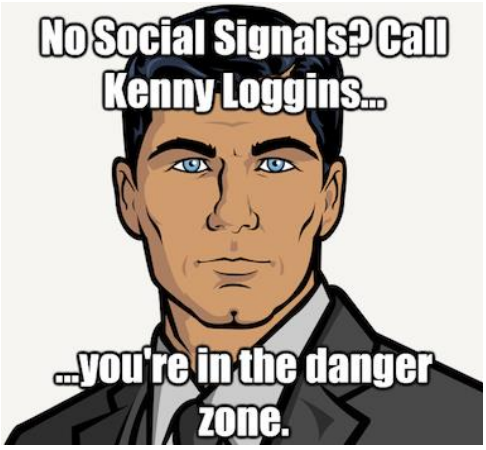
Procedurally, the one-sample t-test and z-test are identical. Incidentally, the procedures in NHST are basically the same regardless of the test used and the only thing that differs is the inferential test. In the z-test and one-sample t-test a sample mean (*M*) is compared to a fixed value representing a population mean (μ). The difference between these tests is that in the z-test the value to which the sample mean is compared is an *actual* population mean, whereas in the one-sample t-test the value to which the sample mean is compared is *assumed* to represent a population mean. For both tests, the difference between the sample mean and the population mean is divided by sampling error. For the z-test this sampling error was measured by the standard error of the mean, whereas in the one-sample t-test, because the population standard deviation is unknown, the standard error is estimated by dividing the estimated standard deviation by the square root of the sample size:

$$\widehat{s}_{\overline{X}} = \frac{\widehat{s}}{\sqrt{n}}$$

This **estimated standard error of the mean** was introduced earlier and is the denominator of the one-sample t-test:

$$t = \frac{\overline{X} - \mu}{\widehat{s}_{\overline{X}}}$$

11.2 One Sample t-test



Say we are media researchers conducting a study on television viewing. We want to know (a) how much time students at a Rhode Island university spend watching the cartoon Archer per week, and (b) whether the amount of time spent watching Archer significantly differs from 10 hours per week. Let's say 10 hours per week of television viewing is assumed by the null hypothesis, but this is not an actual population parameter.

A sample of Rhode Island college students is selected and each student is asked how many hours per week they watch Archer. The mean number hours per week this sample spends watching Archer per week will be calculated and compared to 10 hours per week. We collect data from n = 50 students at Rhode Island universities and find the mean number of hours spent watching Archer is M = 11 hours per week, with an estimated standard deviation of $\widehat{s} = 3$. (We set our alpha-level to α = .05.)

The steps in performing the one-sample t-test are identical to those of the z-test and we begin by stating the hypotheses (as before, I will write out and interpret both a directional and a non-directional version of the hypothesis):

H_0: $\mu_{\text{RI Students}} = 10$	or	H_0: $\mu_{\text{RI Students}} = \mu_{10}$
H_0: $\mu_{\text{RI Students}} > 10$	or	H_0: $\mu_{\text{RI Students}} > \mu_{10}$

Next, calculate the test statistic in the one-sample t-test. Recall, the sample mean for number of hours/week Archer was watched was 11 hours/week and that this mean value was being compared to a predicted mean of 10. First, we calculate the estimated standard error of the mean using the estimated standard deviation ($\hat{?} = 3$) and the sample size ($n = 50$):

$$\hat{s}_{\bar{X}} = \frac{3}{\sqrt{50}} = 0.424$$

This value is the denominator in the one-sample t-test. The outcome of the one-sample t-test is:

$$t = \frac{11-10}{0.424} = 2.358$$

This is the **obtained t-Value** and the *test statistic* used to determine whether the difference between the sample mean and the assumed population mean is statistically significant.

To determine the statistical significance of this outcome, we need to use t-tables, which is Table 2 in Appendix A (Probabilities Under the t-Distribution). Using this table is nearly identical to using Table 1, the standard normal tables, to determine statistical significance; however, when looking up the probability associated with your test statistic you also need to take into account the degrees of freedom.

From our example, the test statistic ($t = 2.358$) rounds to $t = 2.36$ and the degrees of freedom in the sample (50 – 1 = 49) is closest to the column associated with df = 40. (Note, you should never use a df value greater than your actual degrees of freedom.) The column associated with 40 degrees of freedom and the row associated with $t = 2.36$ are highlighted in the table below:

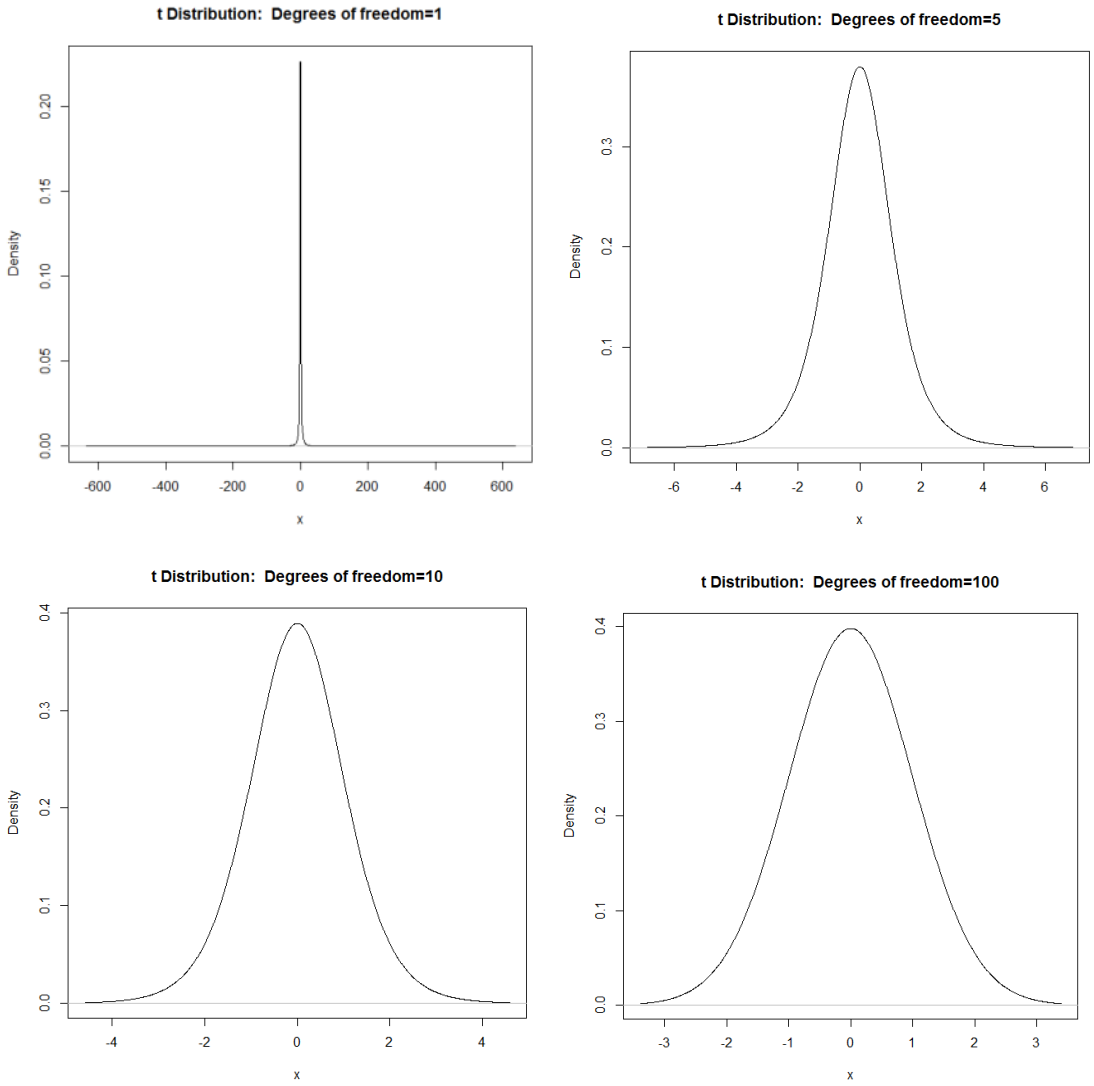
												df					
t	**1**	…	**16**	**17**	**18**	**19**	**20**	**22**	**24**	**26**	**28**	**30**	**40**	**60**	**100**	**200**	**1000**
2.31	.1300	…	.0173	.0169	.0165	.0161	.0158	.0153	.0149	.0145	.0142	.0140	.0131	.0122	.0115	.0110	.0105
2.32	.1295	…	.0169	.0165	.0161	.0158	.0155	.0150	.0146	.0142	.0139	.0137	.0128	.0119	.0112	.0107	.0103
2.33	.1290	…	.0166	.0162	.0158	.0155	.0152	.0147	.0143	.0139	.0136	.0134	.0125	.0116	.0109	.0104	.0100
2.34	.1286	…	.0163	.0159	.0155	.0152	.0149	.0144	.0140	.0136	.0133	.0131	.0122	.0113	.0106	.0101	.0097
2.35	.1281	…	.0160	.0156	.0152	.0149	.0146	.0141	.0137	.0133	.0130	.0128	.0119	.0110	.0104	.0099	.0095
2.36	.1276	…	.0157	.0152	.0149	.0146	.0143	.0138	.0134	.0130	.0127	.0125	.0116	.0108	.0101	.0096	.0092
2.37	.1271	…	.0153	.0149	.0146	.0143	.0140	.0135	.0131	.0127	.0125	.0122	.0113	.0105	.0099	.0094	.0090
2.38	.1266	…	.0150	.0146	.0143	.0140	.0137	.0132	.0128	.0125	.0122	.0119	.0111	.0103	.0096	.0091	.0087
2.39	.1261	…	.0148	.0144	.0140	.0137	.0134	.0129	.0125	.0122	.0119	.0117	.0108	.0100	.0094	.0089	.0085
2.40	.1257	…	.0145	.0141	.0137	.0134	.0131	.0126	.0123	.0119	.0116	.0114	.0106	.0098	.0091	.0087	.0083

From Table 2 you can see this test statistic is associated with a *p*-value of $p = .0116$, which is the *p*-value associated with a directional alternate hypothesis. Because this *p*-value is less than our alpha level ($\alpha = .05$), the sample mean ($M = 11$) is significantly different from 10 hours per week, that is, there is a statistically significant difference between the sample mean and the assumed population mean of 10. Because the difference is statistically significant, we reject the null hypothesis and accept the alternate hypothesis.

Note, the p-values in Table 2 are associated with directional (one-tailed) predictions. If you have a non-directional alternate hypothesis you must look up a *p*-value and then multiply that value by 2 to get the *p*-value for a non-directional prediction. In this case, the *p*-value that we looked up ($p = .0116$), which is associated with a directional alternate hypotheses, is equal to 2 x .0116 = 0.0232 for a non-directional alternate hypotheses.

11.3 t-Distribution Approximation to the Normal Distribution

Previous chapters revealed that larger sample sizes are generally preferred over small samples, because the larger a sample the better it represents a population. Also, as you increase sample size the sampling distribution of means better approximates a normal distribution. This can be seen in **t-distributions** (distributions of critical t-values) for various degrees of freedom. For the t-distribution, when the degrees of freedom are small, the distribution is flatter and spread out. As sample size and degrees of freedom increase the shape of the t-Distribution approximates the shape of a normal distribution:



When degrees of freedom reach infinity, the shape of the t-distribution equal a normal distribution; hence, for very large sample sizes it is okay to use the z-test instead of the one-sample t-test even if the population parameters are unknown, because the t-test results will be nearly identical to the outcome of the z-test.

11.4 Reporting One Sample t-test in Literature

Just like when reporting the outcome of a z-test in the literature there are several items that need to be reported with the one-sample t-test: (1) sample mean, (2) the value the sample mean is being compared to, (3) the obtained t-value, (4) the estimated standard error of the mean, and (5) the *p*-value. To report just the results of the t-test, you would write out the following: $t(49) = 2.358$, $SEM = 0.424$, $p = .0232$ (two tailed), or $t(49) = 2.358$, $SEM = 0.424$, $p = .0116$ (one tailed). Here, the 49 in parentheses are the degrees of

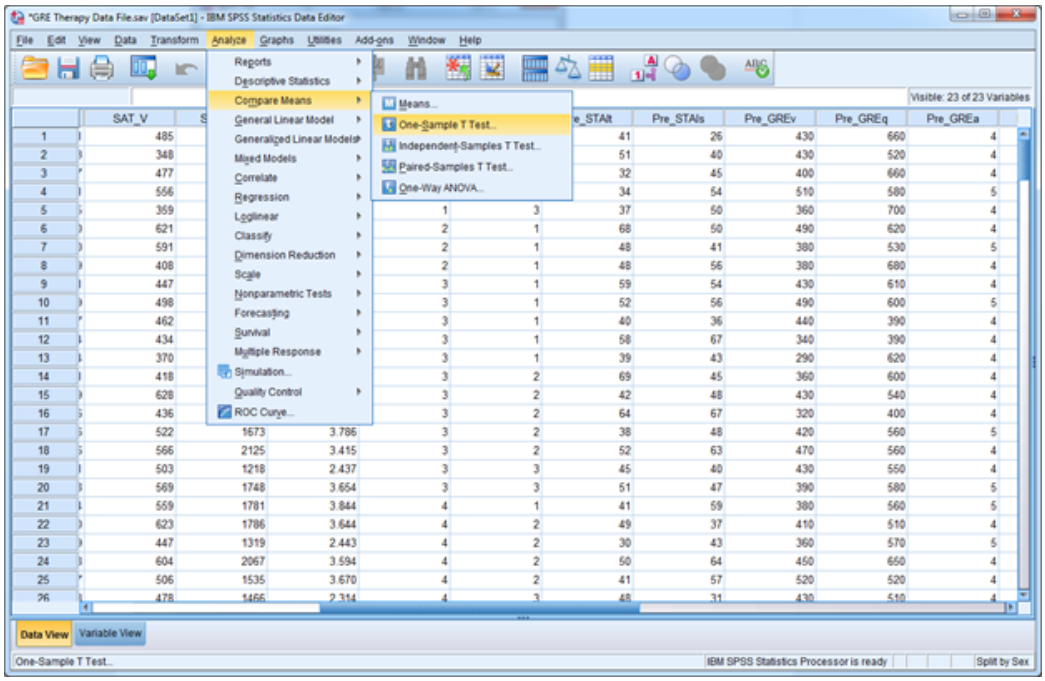
freedom, the SEM stands for the standard error of the mean. Below, I present a generic example of how the results of the Archer example t-test would be written up in a manuscript:

Results

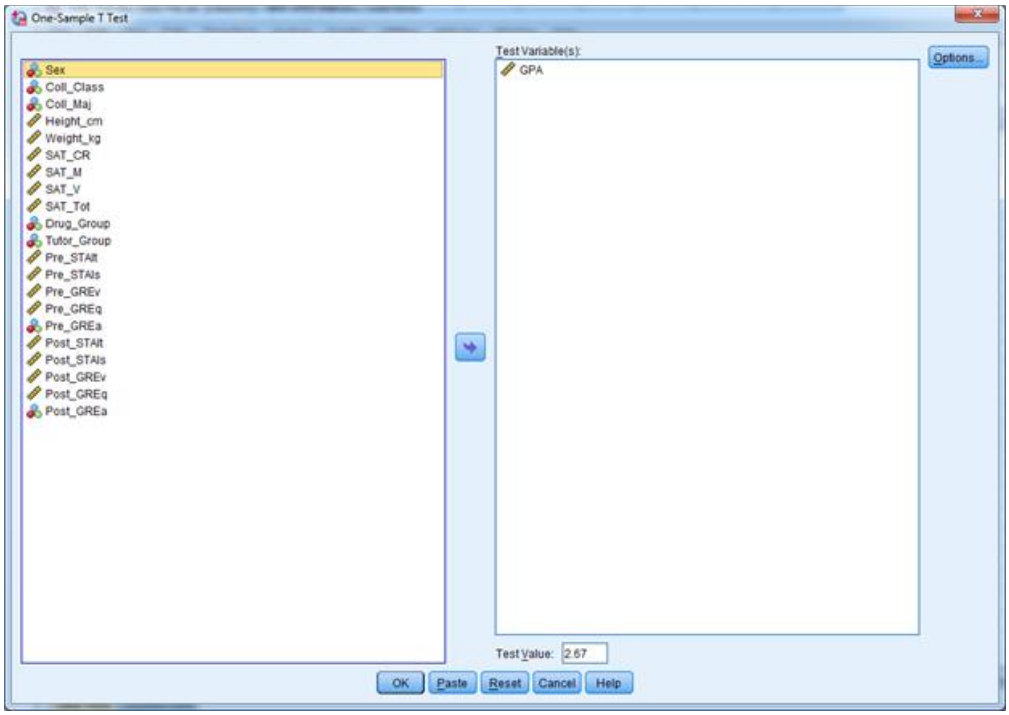
Fifty students from Rhode Island universities were asked to report how much time they spent watching Archer each week. The mean number of hours spent watching Archer per week was found to be *M* = 11 hours/week (*SD* = 3). This sample mean was compared to 10 hours/week. A one-sample t-test yielded a statistically significant difference, *t* = 2.358, *SEM* = 0.424, *p* = .0232 (two tailed).

11.5 One Sample t-test in SPSS

The following uses the GRE Therapy Data file. This data file is based on a hypothetical study examining the influences of a study-aid drug and types of tutoring on performance on the Graduate Record Examinations (GREs). Let's say I want to determine whether the mean GPA of this sample is significantly different from a grade of B-, which is equivalent to a GPA of about 2.67. From the Analyze menu, select Compare Means, and then select One-Sample T Test…



In the new window, select the variable GPA from the list on the left and move it to the panel on the right. This tells SPSS to compare the mean GPA to the Test Value at the bottom. **IMPORTANT**: By default, SPSS sets the Test Value to 0, so you want to compare a mean to a different value you must change it; in this case, to 2.67:



Once you click OK, you get the following output:

T-Test

One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
GPA	240	3.01082	.670476	.043279

One-Sample Test						
	Test Value = 2.67					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
GPA	7.875	239	.000	.340821	.25556	.42608

The first table (one-sample statistics) provides the sample size (N), sample mean (M = 3.01082), standard deviation ($\hat{\sigma}$ = *0.670476*) and estimated standard error of the mean ($\hat{\sigma}_{\bar{x}}$ = *0.043279*). The second table provides the results of the one-sample t-test that compared the sample mean to the test value (2.67). In this example, t = 7.875. The p-value is provided under the column labeled 'Sig. (2-tailed)', which shows p = .000, but means that the actual p-value is less than .001; hence $p < .001$. The Mean Difference is the difference between the sample mean and the test value, and the '95% Confidence Interval of the Difference'

provides the lower (0.25556) and upper limits (0.42608) of the confidence interval around the mean difference (0.340821). (Note, if the t-value, mean difference, and confidence interval values are negative, because the sample mean is less than the test value.) In this example, because $p < .001$ is less than the conventional alpha-level of $\alpha = .05$, we would conclude the sample mean is significantly greater than a grade GPA of 2.67.

CH 11 Homework Questions

1. In what situations would the one sample t-test be used instead of the z-test?

2. State the critical value(s) of *t* for a one-sample *t* test for an alpha level of .05 and of .01 under each of the following conditions:

a. $H_0: \mu = 3, H_1: \mu \neq 3, n = 30$
b. $H_0: \mu = 3, H_1: \mu > 3, n = 30$
c. $H_0: \mu = 3, H_1: \mu \neq 3, n = 50$
d. $H_0: \mu = 3, H_1: \mu < 3, n = 50$
e. $H_0: \mu = 3, H_1: \mu \neq 3, n = 10$
f. $H_0: \mu = 3, H_1: \mu > 3, n = 10$

3. Find the p-value for a one sample t-test under each of the following conditions:

a. $H_0: \mu_1 = \mu_2, H_1: \mu_1 \neq \mu_2, n = 20, t = 2.00$
b. $H_0: \mu_1 = \mu_2, H_1: \mu_1 \neq \mu_2, n = 32, t = 2.50$
c. $H_0: \mu_1 = \mu_2, H_1: \mu_1 < \mu_2, n = 20, t = 2.22$
d. $H_0: \mu_1 = \mu_2, H_1: \mu_1 < \mu_2, n = 10, t = 3.10$
e. $H_0: \mu_1 = \mu_2, H_1: \mu_1 \neq \mu_2, n = 110, t = 2.00$
f. $H_0: \mu_1 = \mu_2, H_1: \mu_1 < \mu_2, n = 50, t = 2.10$

For Exercises 4 – 7, calculate the estimated standard error of the mean, the test-statistic for the given sample mean, find the p-value, and test the viability of the hypotheses using α = .05.

4. Test the variability of the following hypotheses, $H_0: \mu = 100, H_1: \mu \neq 100$; with a sample mean of $\overline{X} = 102$ ($n = 100; \hat{s} = 10$).

5. Test the variability of the following hypotheses, $H_0: \mu = 100, H_1: \mu < 100$; with a sample mean of $\overline{X} = 97$ ($n = 10; \hat{s} = 10$).

6. Test the variability of the following hypotheses, $H_0: \mu = 100, H_1: \mu < 100$; with a sample mean of $\overline{X} = 97$ ($n = 10; \hat{s} = 3$).

7. Test the variability of the following hypotheses, $H_0: \mu = 100, H_1: \mu \neq 100$; with a sample mean of $\overline{X} = 95$ ($n = 25; \hat{s} = 12$).

8. *Use the following information to answer the questions that follow*: The administration of a university wants to know whether the short term memory capacity of college students differs from the short term memory capacity of seven "items" of information. The short term memory capacity for each 16 students is measured. The mean short term memory capacity is found to be $\overline{X} = 8$ items with $\hat{s}^2 = 1.6$.

a. Expressed in terms of μ, what are the null and alternate hypotheses?
b. Calculate the estimated standard error of the mean.
c. What is the obtained t-Value for the sample mean?
d. What is the p-value for this test statistic?

e. Using an alpha level of .01, what decisions should you make about the null and alternate hypotheses? What should you conclude about the short term memory capacity of college students?

9. *Use the following information to answer the questions that follow*: A sample of n = 10 9th grades at James Woods High School can do an average of 11.5 pull-ups (chin-ups) in 30 seconds, with an estimated population standard deviation of $\hat{\sigma} = 3.162$. The US Department of Health and Human Services suggests 9th grades should do a minimum of 9 pull-ups in 30 seconds. Is this sample of 9th grades able to do significantly more pull-ups than the number recommended by the US Department of Health and Human Services? Using this information, answer the following:

a. Expressed in terms of μ, what are the null and alternate hypotheses?
b. Calculate the estimated standard error of the mean.
c. What is the obtained t-Value for the sample mean?
d. What is the p-value for this test statistic?
e. Using an alpha level of .05, what decisions should you make about the null and alternate hypotheses? What should you conclude about the short term memory capacity of college students?

10. *Use the following information to answer the questions that follow*: You want to know whether the average daily temperature for January in Utica, NY differs from 0°C. You measure the daily temperature in Utica, NY for ten random days in January. Here are the recorded temperatures:
-3°C 2°C 0°C 0°C -1°C -3°C -5°C -2°C -2°C -1°C

a. Based on this information, what are the null and alternate hypotheses?
b. Calculate the average daily temperature from these measurements.
c. Calculate the estimated standard deviation.
f. Calculate the estimated standard error of the mean.
g. What is the obtained t-Value for the sample mean?
h. What is the p-value for this test statistic?
i. Using an alpha level of .05, what decisions should you make about the null and alternate hypotheses? What should you conclude about the short term memory capacity of college students?

11. As the degrees of freedom increase, what happens to the t-distribution?

12. Using a word processor (e.g., MS-Word, OpenOffice), write out the results of the t-tests in APA format for #s 4 – 7. Each result should appear on a separate line.

Chapter 12: Bivariate Designs

12.1 What is a Bivariate Design?

"There's a flaw in your experimental design. All the mice are libras."

The comic above illustrates a basic assumption in bivariate designs: that two variables systematically vary with and variation of extraneous variables is controlled. In the comic, the naïve researcher failed to control for the astrological sign of the mice; hence, there is a confounding variable present, which allows alternate explanations for any observable effect of an independent variable on a dependent variable. But, what is a bivariate design?

In both the z-test and one-sample t-test, a sample mean was compared to a value representing a population mean and these two inferential tests are used to determine whether a sample mean significantly differs from a fixed value. But, neither the z-test or one-sample t-test can determine whether a relationship exists between variables. Indeed, it is rare when scientists determine just whether a mean is significantly different from a single value; scientists usually want to test example relationships between variables.

For example, a researcher might manipulate an independent variable by having some students take ginko-baloba while other students take a placebo and measure each student's GPA at the end of the academic year. A researcher could randomly select n = 100 students from the same student population and randomly assign n = 50 students to take a ginko-baloba every day for one school year and the other n = 50 students take a placebo. At the end of the school year grades are collected and compared between the groups (see table below). In this example, a researcher would be examining the relationship between an independent variable (Group: ginko-baloba versus placebo) and the dependent variable (GPA). Whenever a researcher examines the relationship between two variables the researcher has a **bivariate design** and the researcher is examining a **bivariate relationship**.

	Ginko-Baloba	**Placebo**
Mean GPA:	3.30	3.00

Bivariate designs also examine relationships between two dependent variables. An example of bivariate relationship between two dependent variables, height and weight, is below, where you can see that as height increases, weight tends to increase:

Individual	**Height**	**Weight**
A	65”	150 lbs
B	66”	155 lbs
C	68”	160 lbs
D	70”	168 lbs
E	72”	175 lbs

It should be noted, the only way that you can determine whether changes in one variable were potentially *caused* by changes in another variable is by manipulating an independent variable, as in the ginko-baloba vs. placebo example. Assessing the relationship between two dependent variables, as in the height vs. weight example, can only indicate whether the dependent variables are statistically associated.
Bivariate designs that examine the relationship between an independent variable and dependent variable are **experimental designs** or **quasi-experimental designs**. The difference between experimental designs

and quasi-experimental designs is in the independent variable: in experimental designs the independent variable is *manipulated* (e.g., drug versus placebo); whereas in quasi-experimental designs the independent variable is *observed* to change between groups, that is, the levels of the independent variable already designate different groups (e.g., male versus female).

Bivariate designs that examine relationships between two dependent variables are called **correlational designs**. For example, examining the relationship between height and weight, both of which are measurable variables. For each type of design there are different statistical analyses to measure the relationship between the variables and to determine whether the relationship is statistically significant.

12.2 Issues in Experimental Designs

An experimental design is one where an independent variable is manipulated and the objective is to examine this independent variables' influence on some dependent variable. For example, the preceding section described a scenario where some students were given ginko-baloba and other students were given a placebo, and GPAs of all students were then compared between the two groups. Hence, the researcher *manipulated* whether students took ginko-baloba or the placebo.

In an experimental design the group given the treatment condition that is generally of more interest (i.e., the ginko-baloba group in the example above) is the **experimental** (**treatment**) **group**. The group given the non-treatment condition that is of lesser interest (i.e., the placebo group in the example above) is the **control group**. The control group is important, because the behavior of subjects in this group should be normal and unaffected by the treatment level[s] of the independent variable. Behavior and performance of subjects in the experimental group should be different than normal, and this difference should be due to the treatment. Thus, by comparing performance between the control group and the experimental group, one can determine whether the treatment had any effect on the dependent variable. Under the right circumstances a researcher could conclude that taking ginko-baloba *caused* an observed change in GPA. Thus, experiments are used to establish cause-effect relationships between an independent variable and a dependent variable.

But what if the independent variable was not manipulated and the levels of the independent variable were observed to differ between groups (quasi-experimental design)? For example, if a researcher wanted to compare GPAs between males and female students; the levels of the independent variable *sex* are already differs between groups. In most case, the analysis would be exactly the same as that for a true experimental design. That is, the inferential tests that are used to analyze bivariate relationships between an independent variable and dependent variable are the same whether the independent variable was manipulated or not. In quasi-experimental designs, however, you *cannot* make causal statements about any difference observed between the groups. For example, if you were to find a difference in GPA between males and females, you cannot conclude that being male/female *caused* the difference in GPA; all that you can state is there is a difference in GPA between the sexes.

There are a number of issues surrounding the use of experimental designs. Those most relevant to statistics will be described here:

First, in experimental designs a researcher wants to ensure that the groups being compared are identical except for the level of the independent variable. That is, the only difference between the experimental group and control group should be the level of the independent variable. Stated differently, a researcher wants to ensure that subjects in an experimental group have the same characteristics as the subjects in a control group. If all things are equal between the groups, except for the independent variable, any difference in the dependent variable between the groups is most likely due to the independent variable and not to some extraneous variable.

The easiest way to ensure subjects in each group are identical is to place subjects into each group through a process of **random assignment**, where subjects are placed into groups without any discernible order.

Random assignment attempts to eliminate bias toward having certain types of subject in a group; thus, random assignment attempts to equally distribute individual differences between groups. Of course, the only way to know whether random assignment was effective is to measure people on variables other than the dependent variable that may be related to the study. For example, in the ginko-baloba example from the preceding section, you may measure in both the experimental and the control groups each subject's' high school SAT, ACT and IQ. This way, you can examine whether general intelligence levels are equal between the two groups.

In experimental designs you generally want to control as many **extraneous variables** as possible. Extraneous variables are naturally occurring, random variables that are not directly related to a study (e.g., slight changes in temperature, time, lighting, age). Most extraneous variables have no effect on performance or on the outcome of a study; however, it is possible that some variables *could* exert an influence and a researcher may want to control these variables so they are constant across levels of an independent variable. For example, in the ginko-baloba example the amount of time that students study would certainly vary. A researcher may want to set a minimum and a maximum time that students can study each week to exert **experimental control**, that is, hold this extraneous variable constant. Exerting experimental control allows a researcher to conclude with greater confidence that an observed difference in a dependent variable between groups was due to the independent variable, and not some extraneous variable.

It is possible that an extraneous variable may change as the levels of the independent variable change. In the ginko-baloba example, perhaps all the subjects in the ginko-baloba group were psychology majors and all subjects in the placebo group were economics majors. The researcher probably did not intend for this to happen and may not even realize it, but because *college major* varied with the levels of the independent variable, any observed difference in GPA between the control group and the experimental group could be due to the difference in college major. More specifically, say the group taking ginko-baloba (psychology majors) has a mean GPA of 3.30 and the group taking the placebo (economics majors) has a mean GPA of 3.00. What is this difference in GPA due to? If you answered *it could be due to ginko-baloba or to the difference in majors*', or answered '*it cannot be determined*, you're right! Because two things changed as GPA changed you cannot tell whether the difference in GPA can be attributed to giving students ginko-baloba versus a placebo or to the difference in college major. It could be that the psychology major is easy in this particular university or that the economic major is really hard. In such cases where some extraneous variable changes with the independent variable it is a **confounding variable**. Unfortunately, these are difficult to identify and the only correction is to rerun the study without the confounding factor.

Finally, and perhaps most important with an experimental design, is the issue of ***causality***. Remember, if a change in a dependent variable is observed after manipulating an independent variable it may be concluded that changes in the independent variable caused the observed changes in the dependent variable. For causality to be established, three things must occur:

- There must be a manipulation of an independent variable.
- The manipulation of the independent variable must precede an observed change in a dependent variable. This goes without saying. If changes in the dependent variable occur before a manipulation, then changes in the dependent variable may be natural. This is referred to as **temporal precedence**.
- A **systematic relationship** between the manipulated independent variable and changes in the dependent variable must be observed. This also goes without saying. If there is no change in the dependent variable as the independent variable is manipulated, then there is no relationship between the independent and dependent variable.

It is also important to note that an observed change in the dependent variable is caused by changes in the independent variable in the research design, not the statistical test! You can take two groups (children without depression and adults with depression), give them some test (happiness test) and run a statistical analysis. There is nothing saying that you cannot do this, and let's say you find happiness scores of the children are higher (happier) than adults. Does that mean being a child without depression *caused* more happiness? Of course not! Nothing was manipulated! The point is that causality can only be established if

an independent variable is manipulated, which comes through the research design, not the statistical test. Statistical tests cannot tell you whether one thing *causes* another; rather, the tests can only tell you if a relationship is present and is statistically significant.

12.3 Parametric vs. Non-Parametric Statistical Tests

The statistical tests discussed in the following chapters can be classified into one of the following two categories: **parametric statistics** and **non-parametric statistics**. Most inferential statistics are parametric, meaning that the tests make strict assumptions about population parameters and population distributions, usually normal distributions. Parametric tests also make use of means and variances. Two assumptions of parametric statistical tests are the **normality assumption** and the **homogeneity of variance** assumption. The normality assumption means that an inferential statistic assumes the sample data came from a population that was normally distributed; hence, the sample data should be roughly normally distributed. Homogeneity of variance means that an inferential statistic assumes the variance is equal in each different group or condition, such that $\square_1^2 = \square_2^2 = \square_3^2 \ldots \square_\square^2$.

Parametric tests, usually, assume the dependent variable was measured on a ratio or interval scale; ordinal data can be analyzed with parametric tests if the data approximates an interval scale. Nominal scale data generally cannot be analyzed with parametric tests. Examples of parametric tests include the independent groups t-test and the paired samples t-test, which we discuss in the next two chapters.

Non-parametric statistics do not assume the data come from any normally distributed population and there are few, if any, assumptions about a target population. Non-parametric tests do not make use of means and variances; rather, the focus is on the differences between distributions of scores in a sample. Generally, the data comes from nominal or ordinal scales. Some non-parametric tests include the Wilcoxon Rank Sum Test; Mann-Whiney U-test and Chi-Square, some of which we will cover later.

12.4 Between Groups Designs vs. Within-Groups Designs

There are two ways an independent variable can be manipulated: **between-groups/subjects** or **within-groups/subjects**. In a between-groups design (**independent groups design**) each subject is randomly assigned to only one level of the independent variable; thus, each group is independent of the other groups. One of the most important factors in independent groups study is whether assignment of subjects to the conditions/groups was random.

In a within-groups design (**correlated-samples design**, **dependent groups design**, or **repeated measures design**), one group of subjects is exposed to each level of an independent variable; thus, subjects are tested multiple times. Because each subject is tested in each condition there is no need for random assignment, because there is only one group. But selection into that one group should still be random. Importantly, the order in which conditions are presented should be random, or **counterbalanced** across subjects so each permutation of conditions is presented across subjects.

Whether a between-groups design or within-groups design is used depends mainly on the research question; though most scientists prefer within-subjects designs, because the statistical tests are more powerful. Between-groups designs are good if you are conducting a study where you want each subject to be exposed to only one level of the independent variable. For example, in drug studies, because the physiological effects of taking a drug can be long-lasing most researchers do not like to use within-subjects designs, because the effects of one condition can influence performance in a later conditions. In contrast, within-groups designs are good if you are examining performance on some task, but under a variety of similar conditions. In this case, you may want to see how the same people react to different changes in the independent variable.

There are advantages and disadvantages to a within-groups design and between-groups design that are listed in the table below. What is interesting is that the advantage for one type of a design is usually associated to a disadvantage with the other type of design:

	Between-Subjects Design	Within-Subjects Design
Advantages	1. Statistically easier to analyze by hand 2. Each group is independent of the other groups 3. No need to worry about carry-over effects	1. Requires fewer subjects 2. Less costly and less time consuming 3. Statistically more powerful
Disadvantages	1. Time consuming and generally more costly 2. Requires more subjects 3. Less statistically powerful	1. Carry-over and practice effects 2. More difficult to analyze by hand

Carry over effects occur when something you are exposed to influences performance later. Specifically, the influence of a level of an independent variable "carries over" and influences performance at a later point in time. **Practice effects** mean that performance gets better over time (with practice). Thus, facilitation in performance later in an experiment may not be due to the particular condition a person is exposed to; rather, it may be due to performance enhancement through practice.

12.5 Matched Pairs Designs

There is a method not too many researchers take advantage of called a **matched-pairs design**. This is a between-groups experimental design that is analyzed as a within-subjects design. More succinctly, an independent variable is manipulated between-groups, but data are analyzed as though the manipulation was within-subjects. There are two main reasons for this: (1) Within-subjects analyses are more powerful, or (2) the independent variable cannot be *manipulated* within-subjects, but the independent variable is something that can vary within-subjects (e.g., age). In a matched-pairs design the independent variable is manipulated between-subjects. Subjects in each group are ranked in order of overall performance on the dependent variable or performance on another variable. This way the subject with the best performance in the experimental group is "matched" to the subject with the best performance in the control group. The logic being that if this were a within-subjects design an individual with the best performance in the experimental group would likely have the best performance in the control group. Once subjects are "matched," the data is analyzed using the appropriate within-subject analyses.

To use a concrete example, consider the ginko-baloba example. In that study, whether subjects took the ginko-baloba or placebo was manipulated between subjects. A researcher could turn this into a matched-pairs design by rank-ordering the subjects in each group by GPA, from high to low. Once the subjects are ranked the data are then analyzed with a within-groups statistical test.

It is important to note that a researcher usually needs to have some theoretical reason for conducting a matched-pairs design other than a within-subjects design being statistically more powerful. One example is if a researcher wants to study the effect of an independent variable that varied within-subjects, but cannot be manipulated. Such an example would be age: Age changes within a person, but age cannot be manipulated. If a person wants to tests the effects of aging on spatial memory a researcher could sample a group of young people and a group of old people and test spatial memory. A matched-pairs design could be used to contrast spatial memory of young subjects compared to their old subject counterparts.

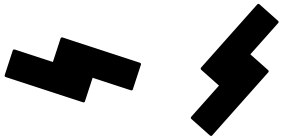
I should also note that there is no specialized statistical test for a matched-pairs design: A matched-pairs design comes from the experimental design itself, not the statistical analysis. Matched-pairs designs use the exact same analyses as within-subjects designs.

CH 12 Homework Questions

1. Describe the differences between experimental research designs and quasi-experimental (observational) research designs.

For each of the studies described in Exercises 2 - 4, identify the independent variable, the dependent variable, and indicate whether the independent variable is between subjects or within subjects.

2. Shepard and Metzler (1972) studied the ability of people to form a visual image of an object in the mind and then mentally rotate the object. Each subject was presented with a display containing two objects side by size, and the angle of rotation between the objects was varied from 0° to 180°, in 15° increments. The response time to decide if the objects were the same or were mirror images of each other was recorded.



3. Harvath (1943) studied the influence of environmental noise on problem-solving. Subejcts were randomly assigned to a noise condition where a buzzer was played in the background, or a quiet condition that did not contain the buzzer. Subjects in each group were given 30 problems, and the number of correct solutions was determined for each subject.

4. Rosenthal and Fode (1963) examined the effect of experimenter bias on the ability of rats to learn to navigate a maze. Subjects were told that they were going to teach a rat to learn to navigate a maze. One group of subjects was told that their rats were genetically smart, “maze bright” rats that should show learning during the first day and performance should increase. A second group of subjects was told their rats were genetically dumb, “maze-dull” rats that should show very little evidence of maze learning. Subjects trained their rats to enter one of two sections of a maze, by rewarding the rat when they entered the correct section. Subjects measured the number of times the rat entered the correct section of the maze each day.

5. What is a control group and what is the purpose of including a control group in a research design?
6. For is random assignment? Why is it important?

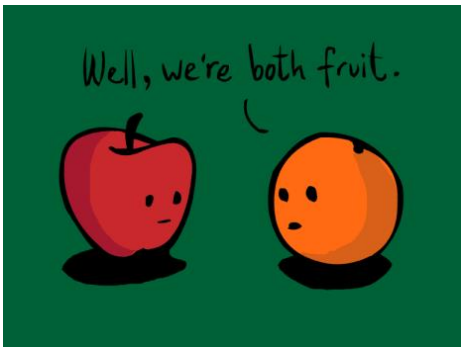
7. What are confounding variables? How can they be controlled?

8. What are the advantages of within-subjects designs compared with between-subjects designs? What is a potential problem?

9. How do parametric and nonparametric statistics differ?

10. What is the normality assumption? What is the homogeneity of variance assumption?

Chapter 13: Independent Groups t-test



13.1 Uses of the Independent Groups t-test

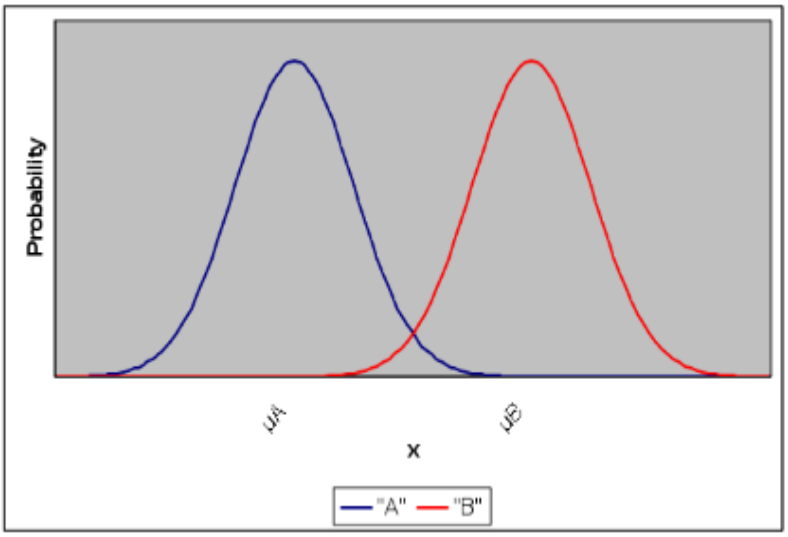
Recall, both the z-test and one-sample t-test compare a sample mean to a value representing a population mean, and in both tests there was no independent variable; hence, performance in a sample was compared to a fixed value. This chapter covers examples where a bivariate test is used when an independent variable is manipulated between groups, and the statistical analysis used in such cases is an **independent-groups t-test**. The independent groups t-test is used when:

1. The dependent variable is *quantitative* in nature
2. The independent variable is *qualitative* in nature, that is, the levels represent different categories.
3. The independent variable has two and only two levels.
4. The independent variable is manipulated between-subjects.

13.2 Mean Differences and Variance between Independent Populations

The independent groups t-test is used to determine whether two sample means, which are assumed to come from two independent populations, are statistically different. In an independent groups design, there are different samples of subjects; for example, a sample of subjects in a drug or experimental group and a sample of subjects in a placebo or control group. It is assumed the subjects in one sample are identical to the subjects in the other sample, except for the level of the independent variable. Subjects in each sample are also assumed to come from different populations, that is, subjects in “Sample A” are assumed to come from “Population A” and subjects in “Sample B” are assumed to come from “Population B.” The populations are assumed to differ only in the level of the independent variable. In short, it is assumed there are two independent samples, each sample coming from its own independent population, with each population differing by only the level of an independent variable. Thus, any difference found between the sample means should also exist between population means, and any difference between the populations means must be due to the difference in the levels of the independent variable.

You can visualize this in the figure below, which presents two populations distributions. Subjects in each population are assumed to be equal except for the level of an independent variable. The distance between the means of the populations is assumed to be due to a difference in the independent variable. Thus, the overlap in the distributions reflects the effect the independent variable on a dependent variable: The more overlap, the less of an effect the independent variable has.



Recall, if you select all possible samples of a specific size (*n*) from a population and calculate the mean of each of sample, you end up with a sampling distribution of the mean. This sampling distribution of the mean approximates a normal distribution as *n* approaches infinity, as per the central limit theorem. The same is true if you have two populations and you select every possible sample of *n* scores from Population A and every possible sample of *n* scores from Population B: you end up with a sampling distribution of means for Population A and a sampling distributions of means for population B. Each of these sampling distributions has a mean equal to the mean of its population (μ_A and μ_B) and has a standard deviation equal to $\sigma\sqrt{n}$. If we assume the sample mean is the best approximation to μ, any difference between sample means should reflect the true difference between population means; hence, a difference in sample means should be due to effect of the independent variable.

Recall, test statistics are ratios of a measure of **effect** to a measure of **error**. The difference between sample means in the scenario above is the effect of the independent variable. So what is the error? Because we have two populations we need to account for the error variance in both populations, that is, the variance in the difference *between* the populations.

Consider this: If you calculated every mean of sample of size (*n*) from a population you end up with a sampling distribution of means, right? If you calculate the difference between every possible pair of sample means, you would end up with a **sampling distribution of mean differences between independent groups**. That is, if you selected every possible sample of *n* scores from Population A and every possible sample of *n* scores from Population B, then took the difference between every possible pair of sample means you end up with a single distribution. This single distribution would be a distribution of *mean differences* between samples of size *n*.

Say that you select a sample from Population A and find it has a sample mean of 10, and you select a sample from Population B and find it has a sample mean of 11. The mean difference is 10 – 11 = -1. If you continue to do this for every possible pair of samples drawn from Population A and Population B, you will find that some differences are greater than zero and some differences are less than zero, but overall the positive differences would cancel out the negative differences. Also, after calculating the differences between sample means, you would find that the resulting sampling distributing of mean differences has a mean difference equal to the true difference between population means ($\mu_A - \mu_B$), which is normally assumed to be zero under the null hypothesis.

The standard deviation of this sampling distribution of mean differences is equal to the **standard error of the mean difference between independent groups**:

$$\sigma_{\overline{X}_1-\overline{X}_2}=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

The subscript 1 and 2 refer to different levels of the independent variable. The standard error of the difference basically combines the variance from each independent population (or level of an independent variable) into one measure of error. Conceptually, the standard error of the difference is the *average difference between sample means of a given size (n) from each population*. Note that the standard error of the difference is actually just the summed standard error of the mean for each distribution:

$$\sigma_{\overline{X}_1-\overline{X}_2}=\sqrt{\frac{\sigma_1^2}{n_1}}+\sqrt{\frac{\sigma_2^2}{n_2}} \quad \text{or} \quad \sigma_{\overline{X}_1-\overline{X}_2}=\frac{\sigma_1}{\sqrt{n_1}}+\frac{\sigma_2}{\sqrt{n_2}}$$

However, this formula is only useful when several conditions are met:

1. The population variances are known, which is almost never the case.
2. The variances in each population are equal, which is almost never the case. Nonetheless, in most circumstances you assume the variances are equal, as per homogeneity of variance.

Because we almost never know the true population variances, you must estimate the standard error of the mean difference between independent groups.

13.3 Pooled Variance and Estimated Standard Error of the Difference

Calculating the estimated standard error of the difference is a two-step process. First, you calculate the **pooled variance estimate**, which is the combined average estimated variance for both populations based on samples of size n_1 and n_2:

$$\widehat{s}^2_{pooled}=\frac{(n_1-1)\widehat{s}_1^2+(n_2-1)\widehat{s}_2^2}{n_1+n_2-2}$$

The subscripts 1 and 2 refer to the two different samples (groups). It does not matter which sample is labeled 1 and which is labeled 2, so long as you are consistent. The pooled variance estimate is the *average estimated variability between the two independent samples*. Because the product of an estimated population variance from a sample and the degrees of freedom of that sample is equal to the sum of squares for a sample, the pooled variance estimate can also be calculated as follows:

$$\widehat{s}^2_{pooled}=\frac{SS_1+SS_2}{n_1+n_2-2}$$

The pooled variance estimate is then used in calculating the **estimated standard error of the difference between independent means**:

$$\widehat{s}_{\overline{X}_1-\overline{X}_2}=\sqrt{\frac{\widehat{s}^2_{pooled}}{n_1}+\frac{\widehat{s}^2_{pooled}}{n_2}}$$

This is the error term for the independent group t-test. It represents the estimated average difference between sample means of size n_1 and n_2 that were selected from independent populations.

13.4 Hypotheses in the Independent Groups Design

Recall, the mean of the sampling distribution of the mean differences between independent samples will have a mean equal to the difference between the population means ($\mu_A - \mu_B$). Usually, this difference is assumed equal to zero, which would indicate there is no effect of an independent variable ($\mu_A - \mu_B = 0$), and any deviation from zero between the sample means is assumed to be due to the independent variable. Thus, under the null hypothesis the difference between means from two independent groups is generally expected to be zero:

$$H_0: \mu_1 - \mu_2 = 0 \quad \text{or} \quad H_0: \mu_1 = \mu_2$$

In short, if under the null hypothesis the two population means are expected to be equal, then the alternative hypothesis is:

$$H_1: \mu_1 - \mu_2 \neq 0 \quad \text{or} \quad H_1: \mu_1 \neq \mu_2$$

Thus, under the alternative hypothesis the difference between the two population means is not expected to be zero. The null and alternate hypotheses above reflect non-directional (two-tailed) predictions, because the alternate hypothesis predicts 'some difference' between the means, but it is also possible to generate directional (one-tailed) predictions. For example, I predict the mean of Population 1 will be greater than the mean of Population 2. The null and alternative hypotheses are:

$$H_0: \mu_1 - \mu_2 = 0 \quad \text{or} \quad H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 - \mu_2 > 0 \quad \text{or} \quad H_1: \mu_1 > \mu_2$$

The alternate hypothesis is predicting the mean of Population 1 will be greater than the mean of Population 2; that is, the difference between the means of Populations 1 and 2 will be greater than zero.

13.5 Degrees of Freedom

Recall, degrees of freedom are equal to $n - 1$ in a sample. In an independent groups t-test, you are comparing <u>two</u> samples and each group has its own $n - 1$ degrees of freedom. For example, say sample A has $n = 10$ people and sample B has $n = 12$ people. In sample A, df = 10 – 1 = 9 and in sample B, df = 12 – 1 = 11. When dealing with independent groups designs, you need to account for the total degrees of freedom across all samples. Most assume the total degrees of freedom are equal to the total number of subjects across both samples minus one; hence, most would guess the total degrees of freedom in this example would be df = 21, because there are 22 subjects. But, in this case, you have only accounted for 21 of the 20 total degrees of freedom from both samples.

Why are there 20 degrees of freedom and not 21? Remember, degrees of freedom are equal to $n - 1$ in a sample. Because there are two samples we need to account for the degrees of freedom in <u>each</u> sample. There are 9 degrees of freedom in sample A and 11 degrees of freedom in sample B, so the total degrees of freedom is df = 9 + 11 = 20. An easier way to get degrees of freedom in an independent groups t-test is $df = n - 2$ where n is the total number of subjects ($n = 22$); hence, df = 22 – 2 = 20.

13.6 Example of Independent Groups t-test

Many colleges and universities require students to take a Freshman Seminar course, where students are acclimated to college life and taught study skills and time-management skills. Let's say at Faber College, freshmen are required to take such a course and this course has always covered basic study skills. One year, a psychologist who conducts research in learning develops a new study technique where students acquire study skills through working in small groups rather than studying alone. This new system should increase GPA relative to covering only basic study skills. The researcher randomly selects $n = 20$ freshmen

and randomly assigns n = 10 to a traditional freshmen seminar course (Basic Study group) and randomly assigns the other n = 10 freshmen into the new group-studying seminar course (Group Study group). All students complete this course and at the end of their freshman year the GPAs from all 20 students are collected and the mean GPA is compared between groups. The researcher predicts that the mean GPA in the Group Study condition will be greater than the mean GPA in the Basic Study condition (directional hypothesis); thus, the hypotheses are:

$$H_0: \mu_{\text{Group Study}} = \mu_{\text{Basic Study}} \qquad H_1: \mu_{\text{Group Study}} > \mu_{\text{Basic Study}}$$

The data (GPAs) for both groups at the end of the semester are presented in the table below. You can see there are n = 10 different students in each condition and each GPA is measured to the nearest hundredth.

Basic Study		Group Study	
Student	**GPA**	**Student**	**GPA**
Bob	2.00	Stew	3.30
Rob	2.30	Roger	3.30
Tom	2.30	Phil	3.30
Judy	2.70	Jen	4.00
Mary	3.00	Christine	3.70
Pete	3.70	Melissa	2.00
Hans	2.70	George	2.00
Pat	3.00	Gina	2.30
Floyd	3.30	Tim	2.70
Marge	2.00	Tony	2.70

The steps for conducting the independent groups t-test, which are going to be similar to those for the paired samples t-test in the next chapter, are as follows:

1. Determine the mean and sum of squares for each sample
2. Using the sums of squares, calculate the estimated variance
3. Estimate the standard error of the mean difference between independent groups
4. Conduct the independent groups t-test
5. Determine the significance of the t-test and make decisions about hypotheses.

First, calculate the means and sums of squares of each group (see table below).

Group Study				**Basic Study**			
Student	**GPA (X_1)**	$(X_1 - M_1)$	$(X_1 - M_1)^2$	**Student**	**GPA (X_2)**	$(X_2 - M_2)$	$(X_2 - M_2)^2$
Stew	3.30	0.37	0.137	Bob	2.00	-0.70	0.49
Roger	3.30	0.37	0.137	Rob	2.30	-0.40	0.16
Phil	3.30	0.37	0.137	Tom	2.30	-0.40	0.16
Jen	4.00	1.07	1.145	Judy	2.70	0	0
Christine	3.70	0.77	0.593	Mary	3.00	0.30	0.09
Melissa	2.00	-0.93	0.865	Pete	3.70	1.00	0.00
George	2.00	-0.93	0.865	Hans	2.70	0	0
Gina	2.30	-0.63	0.397	Pat	3.00	0.30	0.09
Tim	2.70	-0.23	0.053	Floyd	3.30	0.60	0.36
Tony	2.70	-0.23	0.053	Marge	2.00	-0.70	0.49
	$\Sigma X_1 = 29.3$		$SS_1 = 4.382$		$\Sigma X_2 = 20$		$SS_2 = 2.84$
	$M_1 = 2.93$				$M_1 = 2.70$		

Using the sums of squares, we first calculate the pooled variance estimate:

$$\widehat{s}^2_{pooled} = \frac{4.382 + 2.84}{10 + 10 - 2} = 0.401$$

Next, calculate the estimated standard error of the difference between independent groups:

$$\widehat{s}_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{0.401}{10} + \frac{0.401}{10}} = 0.283$$

This is the estimated standard error of the difference for the independent groups t-test; that is, the estimated average mean difference between population means for the Basic Study group and Group Study group. The next step is to perform the independent group t-test and calculate our obtained t-Value:

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\widehat{s}_{\overline{X}1-\overline{X}2}}$$

The numerator includes the actual difference between the two sample means and the hypothesized difference between the two population means. The difference between the two population means is usually assumed to be zero under the null hypothesis. In such cases, the (μ_1 - μ_2) will be equal to zero and can be dropped from the t-test:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\widehat{s}_{\overline{X}1-\overline{X}2}}$$

The next step is inserting the sample means and performing the t-test. This is very important: If you have a non-directional (two-tailed) hypothesis test it does not matter which sample mean that you subtract from the other, because you are predicting "some difference" and that difference could be positive or negative. In contrast, when you have a directional (one-tailed) hypothesis, it does matter which sample mean that you subtract from the other. Follow these rules if you have a directional (one-tailed) hypothesis: (1) If you predict one sample mean to be greater than the other sample mean, then you are predicting a positive difference between sample means and a positive t-Value (test statistic). The mean that you are *predicting* to be greater than the other should be the sample mean with the subscript '1'; that is, the mean that comes first in the numerator. (2) If you predict one sample mean to be less than the other sample mean, and then you are predicting a negative difference between sample means and a negative t-Value (test statistic). The mean that you are *predicting* to be smaller than the other should be the sample mean with the subscript '1'; that is, the mean that comes first in the numerator. Please note that for both of these rules, it does not matter whether this is what you find in the means; the placement of the sample means into the t-test is based on what you predict, not what you observe.

Substituting in our sample means and the estimated standard error of the mean difference calculated earlier, we have:

$$t = \frac{2.93 - 2.7}{0.283} = 0.813$$

This obtained t-Value (t = 0.813) is our test-statistic that we need to look up in Table 2, the t-tables. In this example, there were 18 degrees of freedom and the test statistic (0.813) rounds to 0.81. Using these two values in Table 2, we find that this test statistic is associated with a two-tailed p-value of p = .2170. Because this p-value is greater than the chosen alpha-level (α = .01), we conclude there is insufficient evidence to claim there is a statistically significant difference in GPA between the Basic Study group and the Group Study group, so we retain the null hypothesis and make no decision regarding the alternate hypothesis. In layman's terms, we conclude that exposing freshmen students to a new form of Group Study technique resulted in a small non-significant increase in end of year GPAs compared to students taking a Basic Study skills course.

13.7 Reporting in the literature

Several parameters need to be reported with the independent groups t-test: (a) either both sample means or the mean difference between the sample means, (b) the obtained t-Value, (c) the estimated standard error of the mean difference between independent samples, (d) the total degrees of freedom, and (e) the p-value. Below, I present a generic example of how the results:

Results

Twenty Faber College students were randomly assigned to a Basic Study group (n = 10) or Group Study group (n = 10). At the end of the students' freshman year the GPA of each student was recorded and the mean GPA in the Basic Study group was compared to the mean GPA in the Group Study group. An independent groups t-test revealed a non-significant difference between the mean of the Basic Study group (M = 2.70) and the Group Study group (M = 2.93), $t(18) = 0.813$, $SE = 0.283$, $p = .2170$ (one-tailed).

13.8 Confidence Intervals around the Mean Difference

The margin of error and the confidence intervals can be calculated the difference between the sample means. To do this, you use the standard error of the difference and t_{α} from the independent groups t-test. From the example above, we have (note, because α = .01, this would be the 99% confidence interval around the mean difference):

$$\text{CI}_{.99} = t_{.01}\hat{\sigma}_{\bar{X}_1-\bar{X}_2} = 2.88 \times 0.283 = 0.815$$

And

$$\text{CI}_{.99} = (\bar{X}_1 - \bar{X}_2) \pm t_{.01}\hat{\sigma}_{\bar{X}_1-\bar{X}_2} = (2.93 - 2.7) \pm 2.88 \times 2.88 = 0.23 \pm 0.815$$

Which has lower and upper confidence limits of -0.585 and 1.045. Recall that the null hypothesis predicts that the difference between means will be equal to zero. If zero falls within this confidence interval we can assume that the mean difference is not significantly different from the expected mean difference of zero, which is the case in the example here.

13.9 Effect Size: Eta Squared

One question is how strong an effect the independent variable had on the dependent variable, that is, from the example above is a mean difference of 0.230 a large or a small? To know how much of an effect the independent variable had on the dependent variable, one must calculate a measure of **effect size**. One measure of effect size measure called **eta-squared** (η^2), which is the proportion of variance in the dependent variable that is attributable to the effect of the independent variable. The eta-squared effect size is the ratio of total *treatment variance* (the effect of the independent variable) to total variance; thus:

$$\eta^2 = \frac{SS_{Effect}}{SS_{Total}}$$

The **total sum of squares** ($\mathbf{SS_{Total}}$) is the total variation across all of the scores in a set of data, that is, combined over both levels of an independent variable, what is the total variability among all of the scores? To find the total sum of squares, you need to calculate something called the **grand mean**, which is is the sum of all scores in a set of data divided by the total number of scores while ignoring groups. In the example from Section 14.5 the sum of all the scores (GPAs from both the Basic Study and Group Study conditions) was ΣX = 56.300 and the total number of subjects was n = 20. The grand mean is G = 56.3/20 = 2.815. Note, this is the same value you would get by adding the mean of each condition and dividing by two. The sum of squares total is found by subtracting the grand mean from each score in the data, squaring the differences, and adding the squared differences. This is done in the table below.

The **sum of squares effect** (SS_{Effect}) is calculated by subtracting the grand mean from the sample mean that is associated with each individual. For example, in our example from section 14.5, the student Stew was in the Group Study condition, and that condition had a mean of 2.93. The grand mean (2.815) is subtracted from the mean of Stew's group (2.93). This results in a **treatment effect** of 0.115 for Stew. This is done for each individual in Stew's group, and the same is done for each individual in the Basic Study condition, with the grand mean being subtracted from the mean for that other group (2.7). Thus, Bob and all of the others in the Basic Study condition have a treatment effect of 2.700 - 2.815 = -0.115.

In the table below, each student is listed in the leftmost column, and I have then listed which study group condition each student was in. Each student's GPA is then listed, followed by the mean (*M*) of that student's group. The (X - G) and the $(X - G)^2$ columns show the calculation of the sum of squares total. Finally, the (M - G) and the $(M - G)^2$ columns show the calculation of the sum of squares effect.

				Calculating SS_{Total}		Calculating SS_{Effect}	
Student	**Group**	**GPA (X)**	***M***	**(X - G)**	**$(X - G)^2$**	**(*M* - G)**	**$(M - G)^2$**
Stew	Group Study	3.300	2.93	0.485	0.235	0.115	0.013
Roger	Group Study	3.300	2.93	0.485	0.235	0.115	0.013
Phil	Group Study	3.300	2.93	0.485	0.235	0.115	0.013
Jen	Group Study	4.000	2.93	1.185	1.404	0.115	0.013
Christine	Group Study	3.700	2.93	0.885	0.783	0.115	0.013
Melissa	Group Study	2.000	2.93	-0.815	0.664	0.115	0.013
George	Group Study	2.000	2.93	-0.815	0.664	0.115	0.013
Gina	Group Study	2.300	2.93	-0.515	0.265	0.115	0.013
Tim	Group Study	2.700	2.93	-0.115	0.013	0.115	0.013
Tony	Group Study	2.700	2.93	-0.115	0.013	0.115	0.013
Bob	Basic Study	2.000	2.70	-0.815	0.664	-0.115	0.013
Rob	Basic Study	2.300	2.70	-0.515	0.265	-0.115	0.013
Tom	Basic Study	2.300	2.70	-0.515	0.265	-0.115	0.013
Judy	Basic Study	2.700	2.70	-0.115	0.013	-0.115	0.013
Mary	Basic Study	3.000	2.70	0.185	0.034	-0.115	0.013
Pete	Basic Study	3.700	2.70	0.885	0.783	-0.115	0.013
Hans	Basic Study	2.700	2.70	-0.115	0.013	-0.115	0.013
Pat	Basic Study	3.000	2.70	0.185	0.034	-0.115	0.013
Floyd	Basic Study	3.300	2.70	0.485	0.235	-0.115	0.013
Marge	Basic Study	2.000	2.70	-0.815	0.664	-0.115	0.013
				SS_{Total} = 7.486		**SS_{Effect} = 0.264**	

Using the SS_{Total} and SS_{Effect} values, the eta-squared effect size is:

$$\eta^2 = \frac{0.264}{7.486} = 0.035$$

Only about 3.5% of the variance in the dependent variable (GPA) can be accounted for by the effect of the independent variable. The proportion of variance that cannot be explained is 100% - 3.5% = 96.5%. Thus, about 96.5% of the variance cannot be explained by the effect of the independent variable.

Of course, this process take to find η^2 takes time and there is a simpler method for calculating eta-squared, which makes use of total degrees of freedom and the obtained t-value:

$$\eta^2 = \frac{t^2}{t^2 + df} = \frac{0.813^2}{0.813^2 + 18} = 0.035$$

13.10 Effect Size: Cohen's d

Another index of effect size is **Cohen's *d***. There are several methods for calculating Cohen's d, but the most appropriate is to divide the difference in the sample means by the **estimated pooled standard deviation**, which is just the positive square root of the pooled variance estimate. What's nice about Cohen's d is that it provides a standardized measure that can be used to compare across different studies. Specifically, Cohen's d is a measure of the standardized difference between sample means; that is, it is a distance between two sample means in standard deviations.

From the example above, the estimated pooled standard deviation is:

$$\widehat{s}_{Pooled} = \sqrt{\widehat{s}^2_{Pooled}} = \sqrt{0.401} = 0.633$$

Cohen's d is:

$$d = \left|\frac{\overline{X}_1 - \overline{X}_2}{\widehat{s}_{Pooled}}\right| = \left|\frac{2.93 - 2.7}{0.633}\right| = 0.363$$

Thus, the two sample means are separated by only 0.363 standard deviations. Cohen (1992) provides useful labels to describe how 'big' or how 'strong' of a relationship the effect size indicates. The table below reports the minimum Cohen's d and eta-squared values that correspond to 'weak', 'moderate', and 'strong' effect sizes (also called 'small', 'medium', and 'large' effect sizes):

Effect Size	Cohen's d	eta-squared (η^2)
"Small" ("Weak")	.20	.01
"Medium" ("Moderate")	.50	.06
"Large" ("Strong")	.80	.14

Based on our eta-squared value ($\eta^2 = 0.035$) and or Cohen's d value (d = 0.363), the strength of the effect of the independent variable in the example above would represent a 'weak' or 'small' effect of taking ginko-baloba versus a placebo on GPAs.

13.11 Statistical Power

Recall, from earlier chapters, statistical **power** (**1 – β**) is the probability of correctly rejecting a false null hypothesis. You want this probability to be high so that a statistically significant result likely reflects a true significant difference, that is, the null is probably false. The conventional level of statistical power for an inferential test is .80 or greater (note that power has a limit of .999999....). There are a number of methods that can be used to determine the power of an inferential test. Below, I show you how to estimate the power in an inferential test from a program called G*Power (Faul, Erdfelder, Lang & Buchner, 2007). Following that I show you how to determine the number of subjects needed for a study based on a desired level of power.

A **power analysis** takes one of two main forms. A **post hoc power analysis** is performed after data collection and after an inferential test has been conducted on the collected data to determine the **achieved power** of a result. The achieved power is the probability that the results of your inferential test would correctly reject a false null hypothesis. Obviously, if you want to reject the null, you want to have high achieved power (generally .80 or above).

An **a priori power analysis** is performed prior to any data collection or data analysis to determine the sample size needed to arrive at a level of **desired power**. Researchers—should—start a project by

selecting a level of desired power they want to have after data is collected and inferential tests are performed. That is, say you want to achieve power of .80 after an independent group t-test is performed on a set of data. Based on this desired power and several other parameters you perform a power analysis to determine how many subjects are needed to have that amount power. This is important, because sample size is related to power and sample size is something that a researcher has control over. Below, I list the parameters that are associated with statistical power and their general relationship with power:

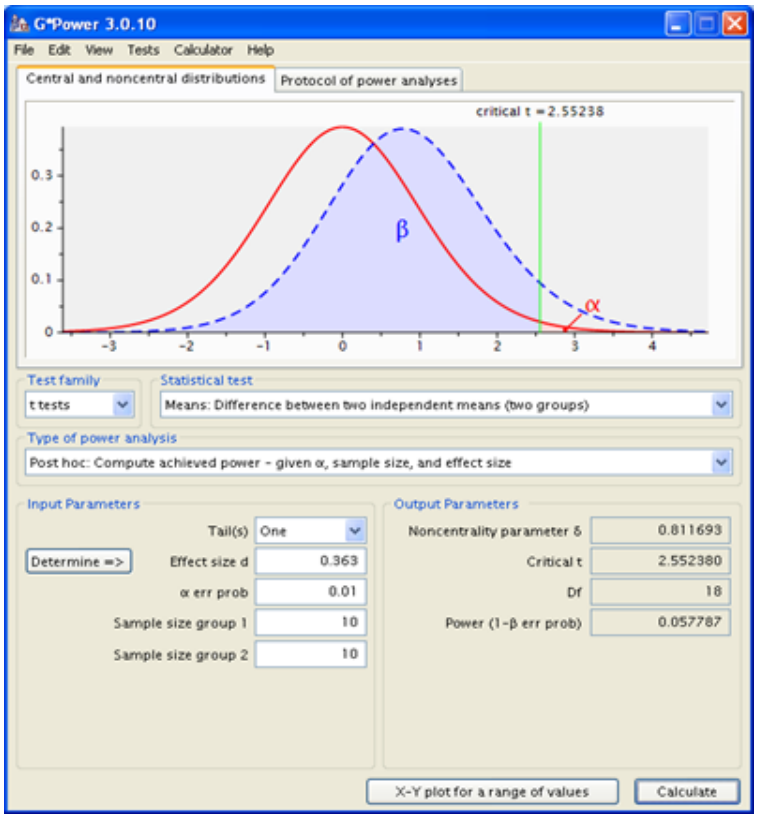
1. **Effect size**: The larger the effect size the more power is in an inferential test.
2. **Alpha level:** The smaller the alpha level, the less power is in an inferential test.
3. **Directionality:** Non-directional hypotheses are less powerful than directional hypotheses.
4. **Sample size and power**: Typically, larger sample sizes have more power.

Importantly, if strength (effect size) of the true relationship between variables is small, it will be difficult to detect the relationship. In this case, larger sample sizes are better and will increase statistical power. In contrast, if the strength (effect size) of the true relationship is strong, it will be easy to detect the relationship, hence, sample sizes do not have to be as large to have sufficient statistical power.

Using the example above, we can calculate the achieved power of the t-test. Recall, the number of subjects per group was n = 10, α = .01, the alternate hypothesis was directional, and the effect size was d = .363. To determine achieved power, use G*Power 3.[1]

Once you open G*Power 3 (and hopefully looked through the Power Packet on the course website): under "Test family" be sure "t tests" is chosen; under "Statistical test" choose "Means: Difference between two independent means (two groups)"; and under "Type of Power Analysis" choose "Post hoc: Compute achieved power." These choices change depending on the type of inferential test you use and the type of power analysis you perform (a priori or post hoc). Next, you'll need to enter the α-level, sample sizes, and effect size from the t-test. The alpha level was α = .01, we had a directional hypothesis, and d = .363.

[1] Go to http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/ and download G*Power 3 for free!

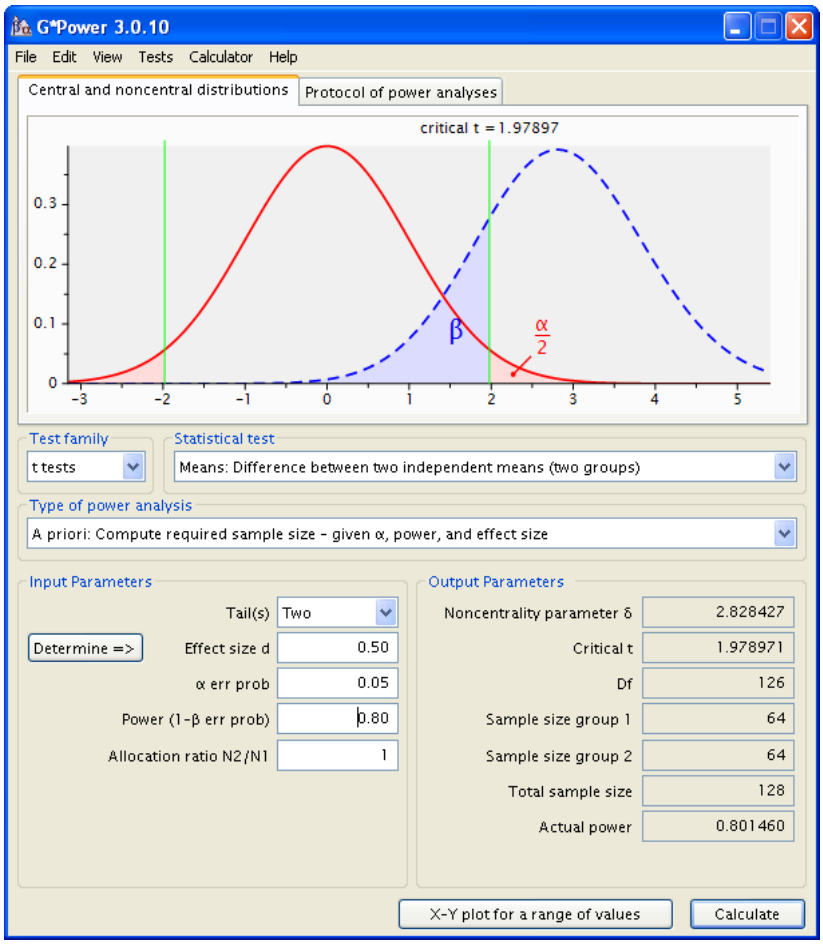


After clicking Calculate you are provided with the achieved power, that is, the probability of correctly rejecting a false null hypothesis. In this case, Power = .057787, which is extremely low. Given we have such low power we cannot conclude with much certainty that the null hypothesis is truly false; we need more power, which requires more subjects. What this could mean is our result is actually a Type I error!

G*Power can also be used to perform an a priori power analysis to determine how many subjects are needed based on a desired level of power. For example, say I plan to run a study that will compare two levels of an independent variable: A_1 vs. A_2. Based on previous research I know the effect size is generally *d* = .50 when these two levels of the independent variable are compared. I plan to use α = .01 and my hypothesis will be non-directional (two-tailed). I want to correctly reject the null hypothesis with Power = 0.80, so I need to use the appropriate number of subjects.

Open G*Power 3. In this example I am running an *a priori* power analysis, so under "Test family" be sure "t tests" is chosen, under "Statistical test" choose "Means: Difference between two independent means (two groups)"; and under "Type of Power Analysis" choose "A priori: compute required sample size." Next, you'll need to enter the α-level, expected effect size, desired power, and choose "two" for Tail(s). Leave the "Allocation ratio N1/N2" set at 1.

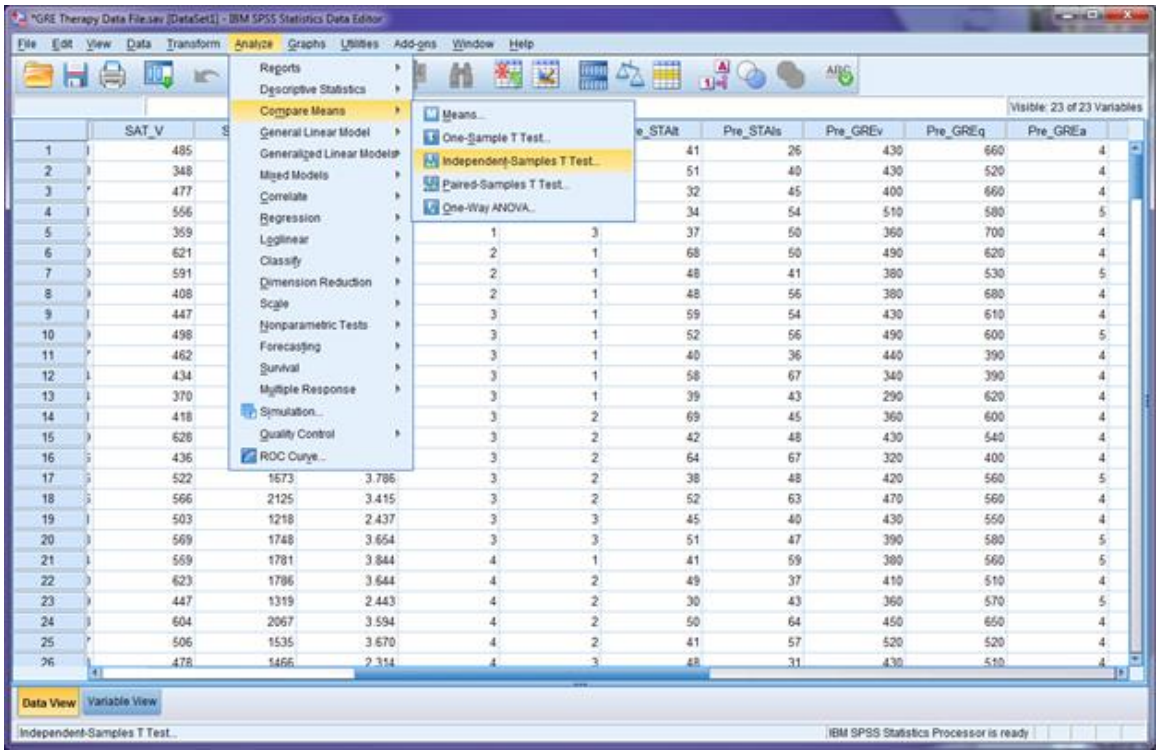
After clicking Calculate, you are provided with the sample size for each group (*n* = 64 for each group), the total sample size (*n* = 128), critical *t*-Value (t = 1.979), and the degrees of freedom (*df* = 126). To correctly reject the null hypothesis in this planned study with probability of .80, I need 128 subjects.



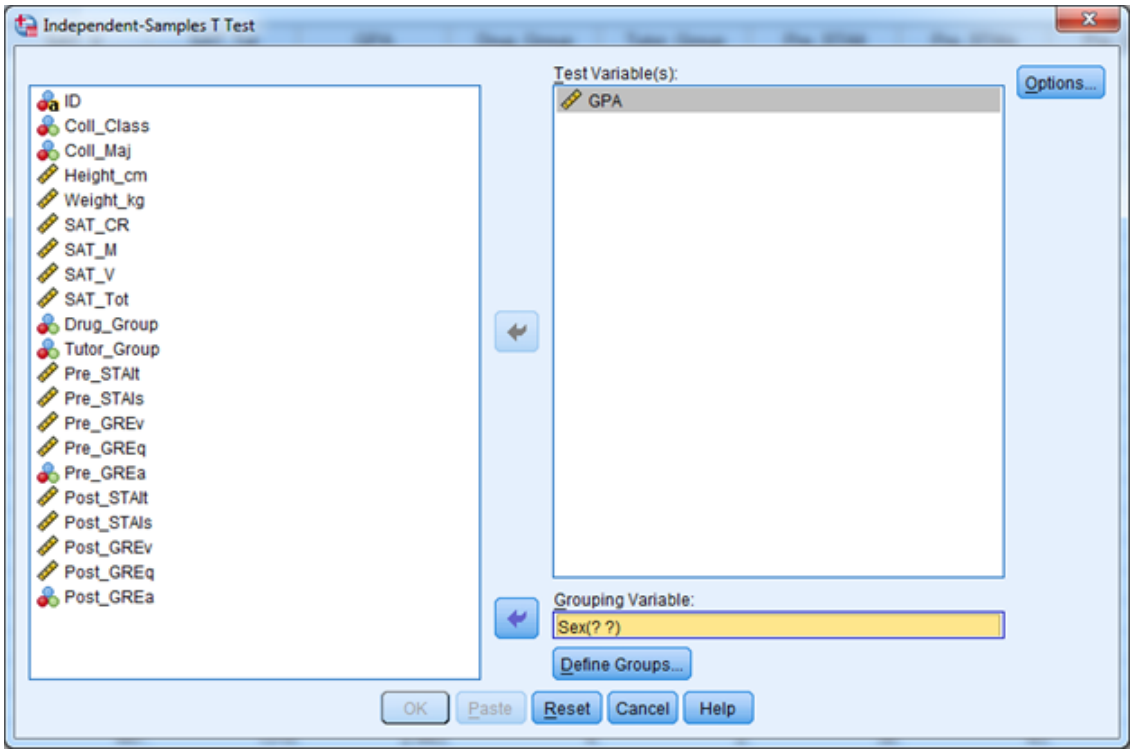
What if the expected effect size was *d* = .80 instead of *d* = .50? How many subjects would I need, if everything else remained the same? In this case, the number of subjects is *n* =26 per group for a total sample size of *n* = 52. Remember, larger expected effect sizes will be easier to detect so fewer subjects are needed.

13.12 Independent Groups t-Test in SPSS

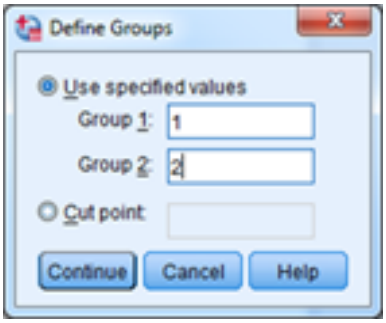
The following uses the GRE Therapy Data file. This data file is based on a hypothetical study examining the influences of a study-aid drug and types of tutoring on performance on the Graduate Record Examinations (GREs). To have SPSS perform an independent groups t-test, from the Analyze menu, select Compare means, and then select Independent-Samples T test:



In the window that opens, you need to place the dependent variable of interest in the Test Variable(s) area and the independent variable in the Grouping Variable area. (Note, you examine differences in more than one dependent variable between the same two groups.) As an example, say we want to compare SAT Math (SAT_M) scores between males and females:



Notice the (? ?) next to the variable Sex. Before you can run the t-test you must tell SPSS which groups you want to compare. In this example, we are comparing the male groups to the female group, and in the data file, males are dummy-coded as 1 and females are dummy-coded as 2. Click on Define Groups and in the popup window enter the dummy-codes for each group (see right).



Click continue and in the main window click OK. You will get two tables in the SPSS output (see below).

T-Test

Group Statistics

	Sex	N	Mean	Std. Deviation	Std. Error Mean
GPA	Males	109	3.04722	.672054	.064371
	Females	131	2.98053	.670226	.058558

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
GPA	Equal variances assumed	.000	.994	.767	238	.444	.066686	.086999	-.104701	.238073
	Equal variances not assumed			.766	229.908	.444	.066686	.087021	-.104775	.238146

The first table (Group Statistics) provided the sample sizes (N), means, standard deviations, and estimated standard errors of the mean for the groups being compared in the t-test. The second table (Independent Samples Test) provides the results of the t-test. Two things to note about the output: First, you can ignore everything under Levene's Test, as this is not relevant for this course. Second, unless noted, you should always equal variances (homogeneity of variance) between groups; hence, you should use the information from the row associated with 'Equal variances assumed'.

In the output, the test statistic is $t = 0.767$ with $df = 238$, and $p = .44$. Because the obtained p-value is above the conventional alpha-level ($\alpha = .05$) the outcome is not statistically significant. The mean difference (0.066686) is the difference between the two mean from the Group Statistics table, and the Std. Error of the Difference is the denominator in the t-test. The values under the 95% Confidence Interval of the Difference are the lower and upper confidence interval limits around the mean difference.

CH 13 Homework Questions

1. When would the independent groups t-test be used to analyze a bivariate relationship?

2. What will the mean of a sampling distribution of the difference between two independent means be equal to?

3. Determine the p-value for an independent groups t-test under each of the following conditions:

a. $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$, $n_1 = 8$, $n_2 = 8$, $t = 3.00$
b. $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$, $n_1 = 10$, $n_2 = 10$, $t = 2.50$
c. $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 > \mu_2$, $n_1 = 9$, $n_2 = 9$, $t = 2.20$
d. $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 < \mu_2$, $n_1 = 25$, $n_2 = 25$, $t = 1.75$
e. $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$, $n_1 = 20$, $n_2 = 20$, $t = 2.10$
f. $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 > \mu_2$, $n_1 = 5$, $n_2 = 5$, $t = 3.10$

4. Determine the critical t-value(s) for an independent groups t-test for an alpha level of .05 and .01 under each of the following conditions:

a. $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$
   $n_1 = 8$, $n_2 = 8$
b. $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 > \mu_2$
   $n_1 = 8$, $n_2 = 8$
c. $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$
   $n_1 = 20$, $n_2 = 15$
d. $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 > \mu_2$
   $n_1 = 20$, $n_2 = 15$
e. $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$
   $n_1 = 8$, $n_2 = 17$
f. $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 < \mu_2$
   $n_1 = 8$, $n_2 = 17$

5. Given the following information, compute the estimated pooled variance and the estimated standard error of the mean difference between independent groups for $n_1 = 10$, $n_2 = 13$, $\widehat{s}_1 = 5.575$, $\widehat{s}_2 = 4.235$.

6. Given the following information, compute the estimated pooled variance and the estimated standard error of the mean difference between independent groups for $n_1 = 50$, $n_2 = 50$, $SS_1 = 40000$, $SS_2 = 60000$.

7. *Use the following information to answer the questions, below*: The following information comes from two samples that were randomly selected from their respective populations:

Sample A	Sample B
$n_A = 100$	$n_B = 100$
$M_A = 120$	$M_B = 130$
$SS_A = 1025$	$SS_A = 2050$

a. Compute the estimated pooled variance.
b. Compute the estimated standard error of the mean difference between independent groups.
c. Compute the test-statistic (t-value) between the means (assume a non-directional hypothesis).
d. Assuming a non-directional alternate hypothesis what is the p-value?
e. What decision should you make about the difference between the two means?

8. *Use the following information to answer the questions, below*: The following information comes from two samples that were randomly selected from their respective populations:

Sample A	Sample B
$n_A = 5$	$n_B = 8$
$M_A = 15$	$M_B = 12$
$SS_A = 125$	$SS_B = 135$

a. Compute the estimated pooled variance.
b. Compute the estimated standard error of the mean difference between independent groups.
c. Compute the test-statistic (t-value) between the means (assume a non-directional hypothesis).
d. Assuming a directional alternate hypothesis what is the p-value?
e. What decision should you make about the difference between the two means?

9. *Use the following information to answer the questions, below*: A professor teaches two sections of the same course and wants to know whether posting audio recordings of his lectures to his website has any effect on student grades. The professor posts audio recordings for one class, and for the other class he posts a written summary of his lecture. Below are the data for both conditions:

Audio Recordings	No Audio Recordings
n = 25	n = 25
$M = 3.35$	$M = 2.55$
SS = 45	$SS_A = 45$

a. Expressed in terms of μ, what are the null and alternate hypotheses?
b. Compute the estimated pooled variance
c. Compute the estimated standard error of the difference between independent groups.
d. Compute the test-statistic (t-value) between the means.
e. What is the p-value?
f. Using an alpha level of .05, what conclusions should you draw about the null and the alternate hypotheses?

10. Use the information from Exercise 9 to answer the following questions:

a. Calculate the eta-squared (η^2). Does this represent a small, medium, or large effect?
b. Calculate the estimated pooled standard deviation.
c. Calculate Cohen's d. What does the value tell you about the separation of the two means?
d. Using G*Power, approximately how much statistical power did this t-test have?

11. *Use the following information to answer the questions, below*:

Group A	Group B
n = 49	n = 49
$M = 10$	$M = 10.7$

$SS = 100$ $SS_A = 120$

a. Compute the estimated pooled variance
b. Compute the estimated standard error of the difference between independent groups.
c. Compute the test-statistic (t-value) between the means.
d. What is the p-value (assuming a directional hypothesis)?
e. Using an alpha level of .05, what conclusions should you draw about the hypotheses?

12. Use the information from Exercise 11 to answer the following questions:

a. Calculate the eta-squared (η^2). Does this represent a small, medium, or large effect?
b. Calculate the estimated pooled standard deviation.
c. Calculate Cohen's d. What does the value of Cohen's d tell you about the separation of the two means?
d. Using G*Power, approximately how much statistical power did this t-test have?

13. *Use the following information to answer the questions, below:* A researcher studied the influence of marijuana on reaction times by having five subjects eat a marijuana laced brownie until a certain level of "stoned-ness" was achieved. Another group ate non-marijuana laced brownies (placebo). All subjects then completed a speeded task. The reaction times (in seconds) are presented in the table:

Marijuana	Placebo
3.00	1.00
3.50	0.50
3.00	1.00
2.00	1.00
2.50	1.50

a. Calculate the means for each group and the sum of squares for each group.
b. Calculate the estimated pooled variance.
c. Calculate the estimated standard error of the difference between independent groups.
f. Compute the test-statistic (t-value) between the means.
g. What is the p-value (assuming a non-directional hypothesis)?
h. Using an alpha level of .05, what conclusions should you draw about the null and the alternate hypotheses?

14. Use the information from Exercise 13 to answer the following questions:

a. Calculate the eta-squared (η^2). Does this represent a small, medium, or large effect?
b. Calculate the estimated pooled standard deviation.
c. Calculate Cohen's d. What does the value of Cohen's d tell you about the separation of the two means?
d. Using G*Power, approximately how much statistical power did this t-test have?

15. *Use the following information to answer the questions, below:* Dr. Jack Daniels examined the effect of alcohol consumption on reaction time. One group of eight subjects consumed alcohol until a certain level of intoxication was reached (determined by physiological measure). A second group of eight subjects was not given any alcohol, but rather, was given a placebo to drink. All participants then completed a reaction time task. The reaction times (in seconds) for each of the two groups are below. Use this information to answer the questions.

Alcohol	Placebo
2.500	1.950
2.250	1.750
2.350	1.350
2.500	2.000
2.600	2.150

2.650	1.750
2.600	1.900
2.550	1.150

a. Assuming the hypothesis is that consuming alcohol will slow reaction times, in terms of μ, what are the null and alternate hypotheses?
b. Compute the mean reaction time for each group.
c. Compute the sum of squares for each group.
d. Compute the estimated pooled variance.
e. Compute the estimated standard error of the difference between independent groups.
f. Compute the test-statistic (t-value) between the means.
g. What is the p-value?

16. For each of the following situations, use G*Power 3 to find the total number of subjects that would be needed to achieve the desired level of Power. (Assume equal group sizes N2/N1 = 1)

a. d = .50; α = .05 (non-directional); Power = .80
b. d = .10; α = .05 (directional); Power = .95
c. d = .80; α = .01 (directional); Power = .85
d. d = .25; α = .01 (non-directional); Power = .90

17. For each of the following situations, use G*Power 3 to find the amount of Power, based on the parameters given. (Assume equal group sizes N2/N1 = 1)

a. d = .25; α = .05 (non-directional); *n* = 76
b. d = .55; α = .01 (non-directional); *n* = 180
c. d = .75; α = .05 (directional); *n* = 30
d. d = .45; α = .01 (directional); *n* = 200

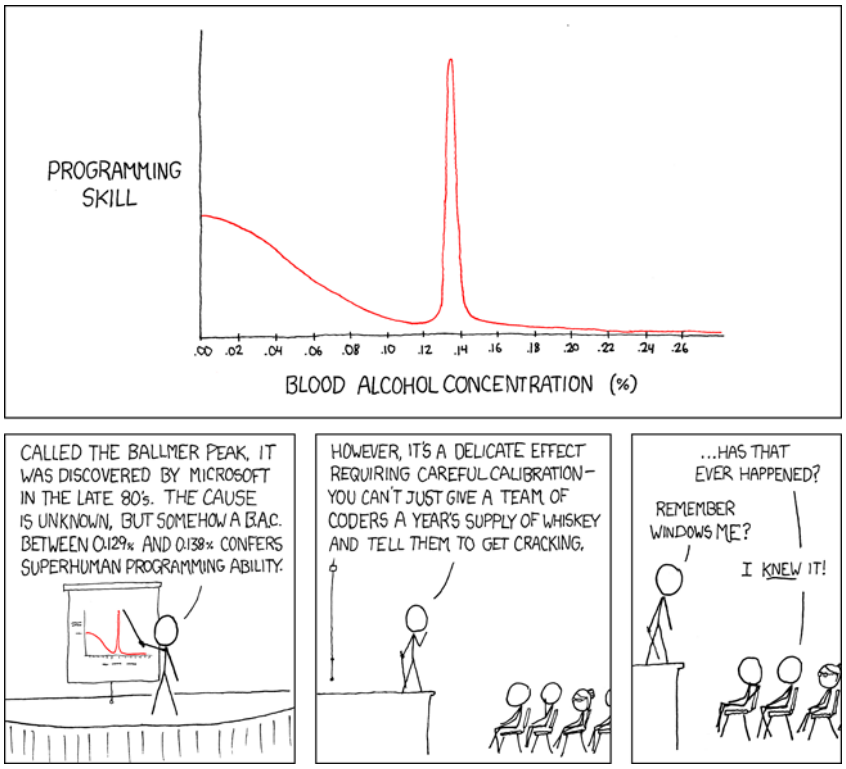
18. Using a word processor (e.g., MS-Word, OpenOffice), write out the results of the t-test performed in #7 in APA format.

19. Using a word processor (e.g., MS-Word, OpenOffice), write out the results of the t-test performed in #8 in APA format.

20. Using a word processor (e.g., MS-Word, OpenOffice), write out the results of the t-test performed in #9 in APA format.

21. Using a word processor (e.g., MS-Word, OpenOffice), write out the results of the t-test performed in #13 in APA format.\

Chapter 15: Correlation



15.1 Relationships between Variables

The comic above provides a good illustration of **correlation**, which is the statistical association between two variables. A correlation exists when changes in one dependent variable are statistically associated with systematic changes in another variable; hence, a correlation is a type of **bivariate relationship**, but one where there is no independent variable. This chapter introduces methods for measuring and describing the strength of the relationship between quantitative variables.

Examples of relationships between variables appear in the table below. Each example lists two variables, X and Y, each with five scores. Think of each x-y pair coming from a single person. Can you identify which examples show a relationship between X and Y and which do not?

Example 1		Example 2		Example 3		Example 4		Example 5	
X	Y	X	Y	X	Y	X	Y	X	Y
2	1	7	1	3	2	2	3	3	2
4	2	6	3	5	4	3	4	4	2
6	3	5	5	7	6	6	3	5	2
8	4	4	7	9	4	4	2	6	2
10	5	3	9	11	2	3	2	7	2

Example 1 shows a relationship between X and Y, because as values of X increase, values of Y increase, so for every change in X there is a consistent change in Y. Example 2 also shows a relationship between X and Y, because as values of X decrease, values of Y increase. Thus, even though scores for the two variables are heading in 'opposite directions', there is still a systematic change in Y that corresponds to the changes in X. Example 3 shows a relationship. Can you see it? Notice as X increases from 3 to 5 to 7, Y

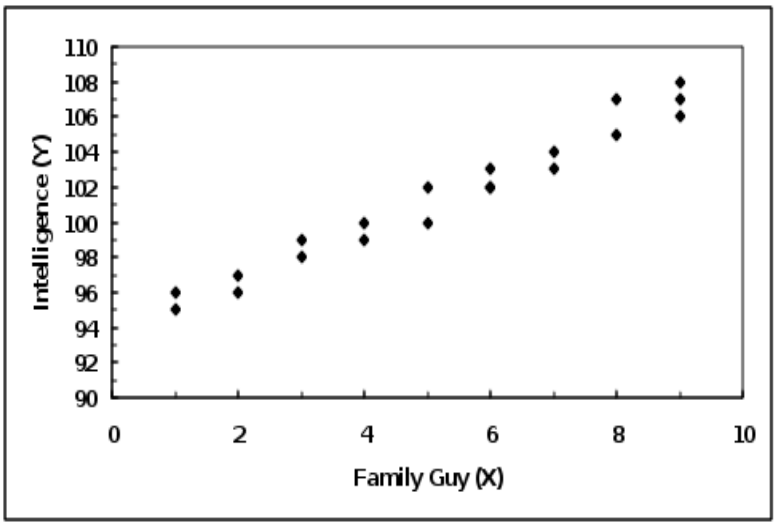
increases from 2 to 4 to 6, but the values of Y begin to decrease from 6 to 4 to 2 as X continues to increase. This is also a relationship, because as values of X increase, there is a systematic change in Y, but then the values of Y decrease. This is a **curvilinear relationship**. Examples 4 and 5 do not show relationships between X and Y. In Example 4, the data are apparently random and there is no consistency in the changes between X and Y. Thus, there is no relationship. In Example 5, as X increases Y does not change, indicating that there are no systematic changes.

15.2 Scatterplots

Determining whether a relationship is present between two dependent variables can be difficult by examining x-y data pairs, because the data is rarely as orderly as presented above. To see if a relationship is present between variables most people begin by creating a **scatterplot**, which is a graph of the x-y pairs for two variables. For example, say we measure *n* = 20 people on the following two variables: (1) *Family Guy Watching*: the number of Family Guy episodes a person watches per week, and (2) Intelligence, as measured by an IQ test. The hypothetical data are presented below:

Person	Family Guy (X)	Intelligence (Y)
A	1	95
B	1	96
C	2	96
D	2	97
E	3	99
F	3	98
G	4	99
H	4	100
I	5	100
J	5	102
K	6	102
L	6	103
M	6	102
N	7	103
O	7	104
P	8	105
Q	8	107
R	9	106
S	9	108
T	9	107

You can see from the data as the number of Family Guy (X) episodes watched increases, Intelligence scores (Y) also increase. In a scatterplot each x-y data point is plotted without lines or bars connecting the data points. Specifically, for the first x-y pair (Person A: x = 1, y = 95) you locate the position on the abscissa for X = 1 and the position on the ordinate for Y = 95, and plot a point at that position in the graph and do the same for the remaining x-y pairs. After you have done this for each x-y pair you have a scatterplot, which can be seen in the figure to the right.

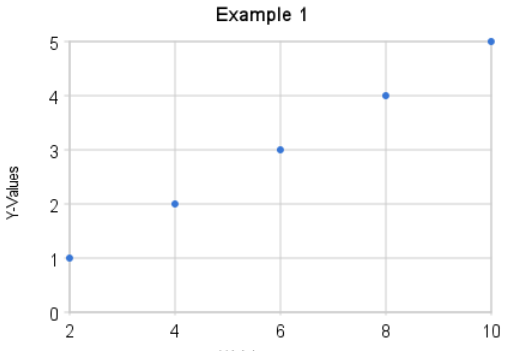


You can see from the scatterplot above, as the X values increase so too

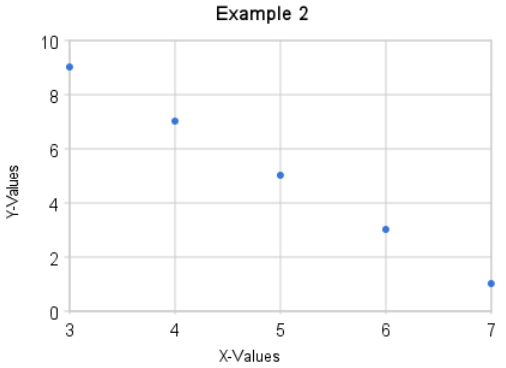
does the Y values. Thus, more Family Guy viewing is associated with higher Intelligence (and less viewing s associated with lower intelligence). It is good to draw a **best fit line** in a scatterplot that shows the trend in the data. Best fit lines should not connect the dots; rather, they just need to show approximately where the relationship is heading.

Scatterplots are used to display a relationship between two variables, and determine whether the relationship is positive linear, negative linear, curvilinear, or absent. A **linear relationship** means that the relationship between variables appears to occur in a straight line; that is, as values of one variable increase the values of a second variable change in a consistent manner. A **curvilinear relationship** exists when there is a change or a "bend" in the relationship between variables. For example, as the values of one variable increase the values of the other variable also increase, but at some point the values of the second variable decrease

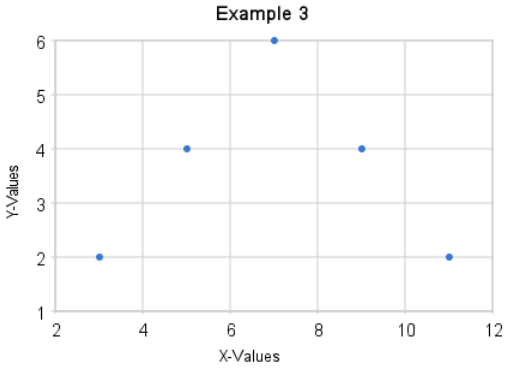
A **positive linear relationship** is observed when the values of both variables have a trend that occurs in the same direction. That is, as the values of one variable increase the values of the second variable also increase. A positive linear relationship can be seen in Example 1 from Section 14.1, which is reproduced in the scatterplot to the right.



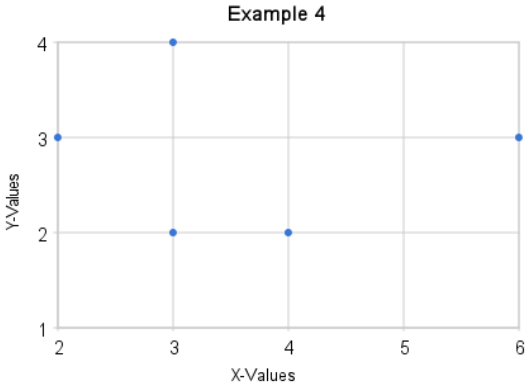
A **negative linear relationship** is observed when the values of variables have trends that occur in opposite directions (*inverse relationship*). That is, as the values of one variable increase the values of the other variable tend to decrease. A negative linear relationship can be seen in Example 2 from Section 14.1, which is reproduced in the scatterplot to the right.

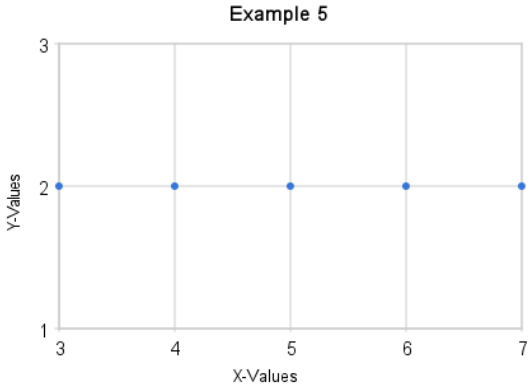


There are many forms of a **curvilinear relationship**, but generally, such a relationship exists whenever there is a change in the relationship between variables. In Example 3 from Section 15.1 as the values of X increased the Y values increased, but then began to decrease. This was a curvilinear relationship. Thus, there is a change in the relationship, which produces a bend (curve) in the scatterplot to the right.



A relationship is ***absent*** whenever there is no systematic change between variables. For example, if the data points in Examples 4 and 5 from Section 15.1 demonstrate cases where no relationship is present. These examples are plotted in the scatterplots below.





15.3 Pearson Correlation Coefficient

The statistic most often used to measures the correlation between quantitative variables is the ***Pearson Product-Moment Correlation Coefficient*** or, more succinctly, the ***Pearson correlation*** (r or r_{XY}). The Pearson correlation is used to measure the *strength* and *direction* of a linear relationship between two quantitative variables. The Pearson correlation can be used to measure the relationship between variables if the following conditions are met:

1. The variables must be *quantitative*; variables cannot be categorical (nominal) or ordinal. Hence, the Pearson correlation can be used only if the values of each variable come from an interval or ratio scale. There is another correlation coefficient, the *Spearman correlation*, which is used to measure the correlation between variables when at least one variable was measured on an ordinal scale. *Chi square analyses* are used if the data comes from nominal scales.

2. Each variable must produce a wide range of its potential values. This is necessary to "capture" the relationship between the variables. If a limited or restricted range of potential values are measured you may not observe the true relationship. For example, say you want to measure the relationship between freshman GPA and SATs. You collect data on $n = 100$ freshmen. You find the range GPAs is 1.00 to 3.80, which is nearly the full range of potential values, but you find that the range of the SATs collected is limited to 1000 to 1200 (the range of potential SAT scores is 600 – 2400). In this case you might not find a relationship between SATs and GPA, because of the restricted range of the SAT scores.

3. The relationship is not curvilinear. The Pearson correlation is designed to measure the direction and degree to which two variables are in a linear relationship. If the relationship between two variables is known to be curvilinear, the Pearson correlation cannot be used.

The Pearson correlation measures the degree to which the relationship between two variables is linear; that is, measures the ***degree of linearity*** between variables. The Pearson correlation has a range between -1.00 to +1.00. The closer to -1.00 or +1.00, the more linear the relationship is between the variables. If the value of the Pearson correlation is found to be equal to -1.00 or +1.00, this indicates that there is a *perfect linear relationship* between variables.

The absolute value of the Pearson correlation is the strength of the relationship between the variables, that is, the degree to which the variables are in a linear relationship. The sign (+/-) of the Pearson correlation tells you whether the relationship is positive-linear or negative-linear; the sign says nothing about the strength of the linear relationship. A perfect Pearson correlation is equal to +1.00 or -1.00. If such a linear relationship is obtained it is possible to predict a value of Y given that we have a value of X, and vice versa. However, it is rarely the case that the relationship between two variables is perfectly linear; thus, it will rarely

be the case that you will obtain a correlation equal to +1.00 or -1.00. In reality, variation of X and Y decreases the strength of linear relationship between variables and the Pearson correlation will deviate from +1.00 or -1.00. A **zero correlation** is a case in which the Pearson correlation is equal to zero ($r = 0$). In this case there is no relationship between variables; they are completely independent. In the next section, I discuss how to calculate the Pearson correlation both in terms of its definition and also computationally.

15.4 Calculating the Pearson Correlation (r)

The Pearson correlation is defined as *standardized covariance between two quantitative variables.* What the heck does this mean? Recall from in Chapter 6 (standard scores) that when calculate a z-Score, you divide the difference between a raw score and the mean of a distribution by the standard deviation of the distribution. This means that the difference between the raw score and the mean was *standardized.* The Pearson correlation is calculated by doing something similar. A measure referred to as *covariance* is divided by the product of two standard deviations; hence, this measure of covariance is standardized. So what is covariance?

Before introducing covariance, let me introduce a set of data that will be used to calculate the Pearson correlation. Below is a set of data for $n = 10$ people. In this hypothetical set of data, assume I measured the age (X) of these ten people and measured the number of books that each of these 10 people read per month (Y). I measured these two variables for these 10 people to see if there is any relationship between a person's age and book reading behavior:

Subject	Age (X)	Books Read Per Month (Y)
A	22	8
B	20	4
C	22	4
D	21	3
E	35	6
F	32	7
G	38	4
H	40	6
I	20	3
J	40	5

Recall, variance is the average variability among scores for a single variable. You can examine the scores for the variable age (X) and for the variable Books Read per Month (Y) in the table to the left and see that those scores vary; this variance. **Covariance** is the *average co-variation of scores between variables*, that is, the average amount by which two variables are changing (varying) together. The difference between variance and covariance is that covariance is the average variation between scores from two variables; whereas variance is the average variation among scores of a single variable. How is covariance calculated? Recall the formula for the estimated variance:

$$\widehat{s}^2 = \frac{\sum(X-\overline{X})^2}{n-1}$$

The formula for **estimated covariance** is similar:

$$cov = \frac{\sum[(X-\overline{X})(Y-\overline{Y})]}{n-1}$$

To calculate covariance you divide the numerator by $n - 1$, rather than by n, because we are estimating the covariance from sample data. Covariance is formally defined as the *average sum of the cross produces between two variables*. The **sum of the cross products** (**SCP**) is the sum of the products of the mean

centered scores from each variable; and this is the numerator in the formula above. That is, the formula for the sum of the cross products is:

$$SCP = \sum[(X - \overline{X})(Y - \overline{Y})]$$

Calculating sum of cross products is similar to calculating sum of squares. Indeed, sum of cross products is conceptually similar to sum of squares, because sum of squares is total variation within a set of scores, and sum of the cross products is the total variation between two sets of scores. The sum of the cross products for the data above is calculated in the table below from the data presented earlier:

Name	Age (X)	Books (Y)	(X - M_X)	(Y – M_Y)	(X - M_X) (Y – M_Y)
A	22	8	-7	3	-21
B	20	4	-9	-1	9
C	22	4	-7	-1	7
D	21	3	-8	-2	16
E	35	6	6	1	6
F	32	7	3	2	6
G	38	4	9	-1	-9
H	40	6	11	1	11
I	20	3	-9	-2	18
J	40	5	11	0	0
	∑X = 290 M_X = 29	∑Y = 50 M_Y = 5			SCP = 43

As you can see in the table above, the first step to calculating the sum of the cross products is to calculate the mean of each variable and subtract the mean from each score. Next, the mean-centered scores for each variable are multiplied to create *cross products*. Finally, the cross-products are summed to obtain the sum of the cross products (SCP = 43). Importantly, unlike the sum of squares, which is always positive, the sum of cross products can be positive or negative: If SCP is positive, covariance will be positive and the correlation is positive-linear. If SCP is negative, covariance will be negative and the correlation is negative-linear. If SCP = 0, it indicates a zero relationship.

The next step is to calculate the covariance. Conceptually, all that you need to calculate covariance is to divide SCP from above by $n - 1$:

$$cov = \frac{SCP}{n-1} = \frac{43}{10-1} = 4.778$$

The next step to calculate a Pearson correlation is to 'standardize the covariance'. The formula for the Pearson correlation is:

$$r = \frac{cov}{\widehat{s}_X \widehat{s}_Y}$$

The numerator is covariance (cov_{xy} = 4.778). The denominator is the product of the estimated standard deviation of each variable. Thus, we must calculate the estimated standard deviation for each variable, which is started by calculating the sum of squares for each variable. This is shown in the table below:

Name	Age (X)	Books (Y)	(X - M_X)	(Y – M_Y)	(X - M_X)(Y – M_Y)	(X - M_X)^2	(Y – M_Y)^2
A	22	8	-7	3	-21	49	9
B	20	4	-9	-1	9	81	1
C	22	4	-7	-1	7	49	1
D	21	3	-8	-2	16	64	4
E	35	6	6	1	6	36	1
F	32	7	3	2	6	9	4
G	38	4	9	-1	-9	81	1
H	40	6	11	1	11	121	1
I	20	3	-9	-2	18	36	4

J	40	5	11	0	0	121	0
	$\sum X = 290$ $M_X = 29$	$\sum Y = 50$ $M_Y = 5$			SCP = 43	$SS_X = 647$	$SS_Y = 26$

The estimated of the standard deviation for each variable are:

$$\hat{s}_X = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{647}{10-1}} = 8.479$$

$$\hat{s}_Y = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{26}{10-1}} = 1.7$$

Now, we have all three pieces needed to calculate the Pearson correlation:

$$r = \frac{4.778}{(8.479)(1.7)} = 0.331$$

It is customary to report the Pearson correlation rounded to two decimal places in APA format; thus, our Pearson correlation would be reported as *r* = .33 in a research report.

One question is what a correlation coefficient of this size indicates. Generally speaking, the larger the absolute value of the Pearson correlation the better. That is, the closer to +1.00 or to -1.00 (the farther away from zero), the stronger the linear relationship between the variables. Nonetheless, a correlation of r = .331 is generally large in the behavioral sciences, because there is so much variation in behavior. Later in the chapter, we'll discuss the effect size of the Pearson correlation.

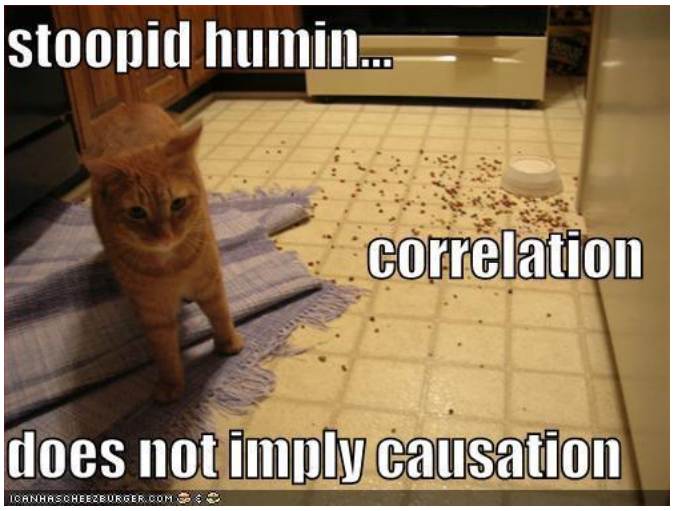
15.5 Proportion of Explained Variance and Residual Variance

The **coefficient of determination** ($\mathbf{r^2}$) is calculated by squaring the Pearson correlation. For the Age-Book Reading relationship above, the coefficient of determination is $r^2 = 0.331^2 = 0.11$. This value is the *proportion of variance accounted for (explained) in the relationship* between the two variables. Specifically, the coefficient of determination is the proportion of x-y pairs in which the values of x and y co-vary in the direction indicated by the Pearson correlation. In the case of a positive linear relationship, the coefficient of determination is the proportion of X scores that increase with the Y scores. In the case of a negative linear relationship, the coefficient of determination is the proportion of X scores that decrease as Y increases. You can think of r^2 as the proportion of scores that are correctly predicted by the relationship. Importantly, the coefficient of determination can <u>never</u> be negative and has a range of 0 to 1, because it is a proportion; bigger values are better.

Residual variance is calculated by subtracting the coefficient of determination (r^2) from 1 ($1 - r^2$). Residual variance is the proportion of variance between two variables that is <u>not</u> accounted for in the relationship; thus, it's the proportion of X-Y scores that do not co-vary together in the direction indicated by the relationship. From the example used here, residual variance = 1 - 0.11 = 0.89. You can consider this to be the proportion of variance that is not accounted for in the relationship.

15.6 Characteristics of the Pearson Correlation

There are several important characteristics you should know about the Pearson correlation. First, a correlation between two variables does not imply that one variable caused the variation in the other variable, that is, *correlation does not mean causation*! From the example in the preceding sections, the Pearson correlation of $r = 0.331$ tells us there is a positive linear statistical association between age and books read per month. The Pearson correlation does not say anything about age *causing* a person to read more. All the Pearson correlation says is two variables are statistically associated. The only way to determine whether changes in one variable cause changes in another variable is by manipulating an independent variable and conducting an experiment.



Second, and related to the preceding point, there is often a misuse and abuse of correlations in the media. Specifically, a relationship between variables can exist, but there may be no causal explanation. That is, there may be a correlation present but manipulations of one variable (in an experimental design) may not have produced the changes in the other variable. For example, there is a known positive correlation between exposure to violent television and the amount of aggression in children, where more violent television exposure is associated with more aggressive behavior in children. It is impossible to tell whether one variable causes the changes in other. It could be that watching violent television causes a child to act aggressive; or it could be that aggressive children like to watch violent television. What is more is that there could be a third variable that *mediates* the relationship; that is, the relationship between violent television and aggression may depend on the presence of some other factor. For example, perhaps this relationship between violent television and aggression exists only in children with conduct disorders and absent in children without the disorder.

Third, occasionally a relationship is found between two seemingly unrelated variables and the relationship is *spurious* (random). Thus, a relationship exists but does not make any sense and occurs without theory; or there is some very simple explanation. For example, that there is a strong positive correlation between number of churches and number of prostitutes: As the number of prostitutes increases the number of churches increases. The relationship exists because of population density. For example, what areas are naturally going to have more prostitutes and more churches? A city! And what areas are likely to have fewer churches and fewer prostitutes? Small/rural towns! Thus, the relationship between churches and prostitutes exists, but for no real meaningful reason: It exists due to natural population density.

Finally, the restricted range problem alluded to earlier occurs when the range of collected scores for some variable represents only a small proportion of all possible scores for that variable. For example, in the memory and cognition literature there is something called '*working memory*' which can be measured with something called the OSPAN test, which ranges from 0 to 1. If you measured the correlation between OSPAN and GPA, but your OSPAN scores were only in the range from 0.50 to 0.60, you have a restricted range of all of the possible OSPAN scores. Thus, you may not detect the true relationship between OSPAN and GPA, only part of the relationship. This is a big problem, because you may find a significant linear relationship if you have a restricted range, when in reality, the relationship between the two variables is curvilinear. Because of your restricted range, you could not detect that relationship.

15.7 Hypotheses and Sampling Distribution of Pearson Correlation

Earlier sections of this chapter addressed how statistical associations between variables can be measured with the Pearson correlation. This section addresses how statistical significance of the Pearson correlation is determined. Determining the statistical significance of a Pearson correlation depends on whether the correlation under the null hypothesis is assumed to be zero or some non-zero value. When the correlation under the null hypothesis is assumed to be zero, you use a type of *t*-test to assess statistical significance. In contrast, when under the null hypothesis the Pearson correlation is assumed to be a value other than zero, use Fisher's z-test to assess statistical significance.

Recall, the value of a correlation has a range of -1.00 to +1.00 and a zero-correlation ($r = 0$) indicates no association between variables. Generally, the null hypothesis predicts no relationship between the variables. The symbol for the Pearson correlation in a population is the Greek lowercase *rho* (ρ). Thus, the null and alternate hypotheses predict that:

$$H_0: \rho = 0 \qquad H_1: \rho \neq 0$$

This is a non-directional alternate hypothesis for a Pearson correlation, because the alternate is stating that the Pearson correlation will be something other than zero. But, the alternate hypothesis is not saying whether the correlation will be positive-linear or negative-linear. It is also possible to generate directional alternate hypotheses for the Pearson correlation. If the alternate hypothesis predicts the correlation will be positive-linear the hypotheses are:

$$H_0: \rho = 0 \qquad H_1: \rho > 0$$

If the alternate hypothesis predicts the correlation will be negative-linear, the hypotheses are:

$$H_0: \rho = 0 \qquad H_1: \rho < 0$$

15.8 Determining Statistical Significance of the Pearson Correlation

Say you are interested in whether the Pearson correlation between age and book reading behavior from the earlier sections is statistically significant. You are not sure whether there is a positive or negative correlation between these variables; thus, you decide to use a non-directional alternate hypothesis:

$$H_0: \rho_{Age,Books} = 0 \qquad H_1: \rho_{Age,Books} \neq 0$$

Recall, from that example, n = 10 and r = 0.331. To determine whether this correlation is statistically significant, we use the following t-test:

$$t = \frac{r}{\sqrt{\frac{(1-r^2)}{(n-2)}}}$$

Before going through the calculations, note this *t*-test is used only when ρ is predicted to be zero under the null hypothesis. First, we'll select an alpha of α = .05 for a non-directional alternate hypothesis. Second, for the Pearson correlation degrees of freedom are equal to $n - 2$, because we need to account for the degrees of freedom in each dependent variable. In the present example, df = 10 – 2 = 8. Plugging *r* and n into the t-test formula above, the obtained t-value is solved below:

$$t = \frac{0.331}{\sqrt{\frac{(1-0.331^2)}{(10-2)}}} = 0.934$$

This is the test statistic that we use to assess the statistical significance of the Pearson correlation. To determine whether this correlation is statistically significant, use Table 2 in Appendix A (t-Tables) by looking up the value of the test statistic (use $t = 0.80$) and degrees of freedom in the correlation (df = 8) to find the *p*-value associated with this correlation, which is $p = .4468$. Because this p-value is greater than the selected alpha level (.05), we conclude the Pearson correlation ($r = 0.331$) is not statistically significant. Thus, we retain the null hypothesis and make no decision about the alternate hypothesis. In layman's terms, we conclude there is insufficient evidence to identify a statistically significant relationship between age and reading.

15.9 Determining Significance when $\rho \neq 0$

In cases where ρ is expected to be something other than zero under the null hypothesis, the *t*-test for the Pearson correlation is inappropriate. This is because the sampling distribution of Pearson correlation coefficients is skewed when $\rho \neq 0$. To overcome this problem, you must make use of the logarithmic transformation of *r* into *r'* (r-prime) and the formula for converting any given *r* into *r'* is:

$$r' = .50[log_e(1+r) - log_e(1-r)]$$

In the formula, *r* is the Pearson correlation and $\log_e$ is the natural logarithm of a number. You could use this formula to determine *r'* or you could use Fisher's z-Transformation Index, which is Table 6 in Appendix A. To use this table, look up the value of your Pearson correlation value (*r*) in the **r(ρ)** columns. The value to the right of this value in the **r'(ρ')** column is the corresponding *r'* value. Importantly, *r'* values are normally distributed, so we do not have to worry about skew.

To test whether a Pearson correlation (r) is statistically different from some non-zero population correlation parameter (*ρ*), You calculate a z-Score for the difference between *r'* and the Fisher's transformed value of *ρ* (*ρ'*) under the null hypothesis, using the standard error of *r'* as the denominator. This is **Fisher's z-test**:

$$z = \frac{r' - \rho'}{\sigma_{r'}}$$

The **standard error of *r'***, which is the error term (denominator) in the formula above, is calculated from:

$$\sigma_{r'} = \frac{1}{\sqrt{n-3}}$$

For example, say the true (population) correlation between studying and sexual activity is ρ = .20. We want to know whether the correlation between studying and sexual behavior for a sample of students from a particular university differs from ρ = .20. Hence, we have a non-directional alternate hypothesis:

H_0: $\rho_{Studying,Sex} = 0.20$ H_1: $\rho_{Studying,Sex} \neq 0.20$

We randomly select $n = 50$ students from this particular university and survey each student to assess the amount of time each student spends studying and the amount of sexual activity each student engages. After measuring studying time and sexual activity for these 50 students we find $r = 0.50$.

To determine the statistical significance of this Pearson correlation, first transform the predicted Pearson correlation under the null hypothesis (ρ = .20) into ρ' using Table 6 in Appendix A. Fisher's transformed value for ρ = .20 is ρ' = .203. Next, transform the obtained Pearson correlation (r =0.50) into an *r*' value using the same table. Fisher's transformed *r'* for $r = .50$ is $r' = .549$.

Calculate the standard error of *r'*:

$$\sigma_{r'} = \frac{1}{\sqrt{50-3}} = 0.146$$

The obtained z-Score is:

$$z = \frac{0.549 - 0.203}{0.146} = 2.37$$

Locate the test statistic (z = 2.37) in column 1 of Table 1 in Appendix A (z-Tables) and locate the probability in column 3 of that row, which is p = .0089 and multiply that value by two for the non-directional *test* to get p = .0178. Because the p-value (p = .0178) is less than the alpha-level of .05, we conclude the correlation between studying and sexual activity (r = 0.50) is significantly different from the population correlation of ρ = .20. Or more generally, this sample correlation between studying and sexual activity is statistically significant.

15.10 Power and Effect size for Pearson Correlation

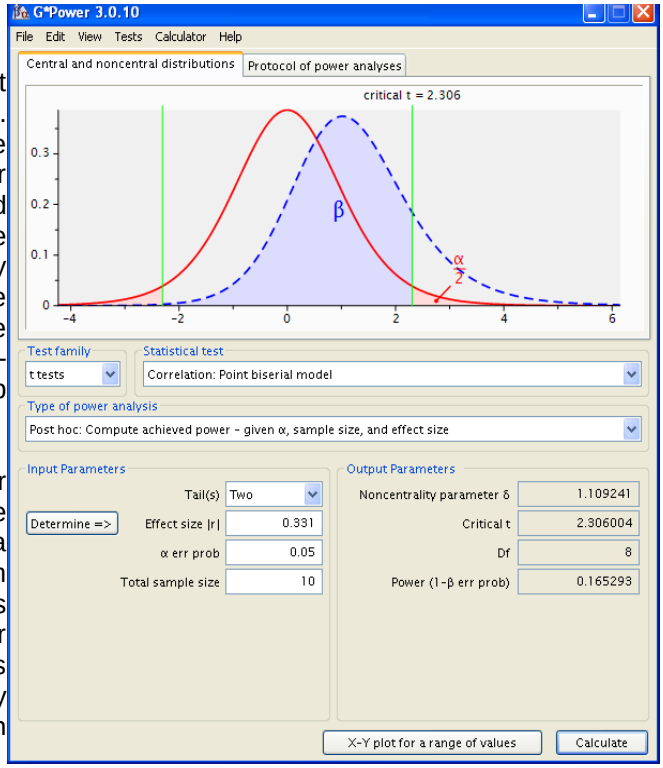
The effect size of the Pearson correlation is the absolute value of the Pearson correlation. That is, the value of the Pearson correlation as a range from -1.00 to +1.00, where values closer to |1.00| indicate a stronger linear relationship. Cohen provides useful labels to describe how strong a relationship is based on the size of a Pearson correlation. The table reports the minimum r and r^2 values that correspond to 'weak', 'moderate', and 'strong' effect sizes for the Pearson correlation (also called 'small', 'medium', and 'large' effect sizes).

Effect Size	**Pearson's r**	**Coefficient of Determination (r^2)**
"Small" ("Weak")	.10	.01
"Medium" ("Moderate")	.30	.09
"Large" ("Strong")	.50	.25

Like with the independent groups *t*-test in Chapter 13 you can use effect size of the Pearson correlation and sample sizes to determine the achieved statistical power (1 - β) of the correlation. G*Power (Faul, et al., 2007) provides a useful tool for performing power analyses on the Pearson correlation. We'll perform a power analysis on the correlation between age and books read per month (r = .331).

Open G*Power 3 and under "Test family," select "t tests" is chosen. Under "Statistical test" choose "Correlation: point biserial method", and under "Type of Power Analysis" choose "Post hoc: Compute achieved power." Next, you'll need to enter the α-level, sample sizes, and effect size (r) from the correlation test above. The alpha level was α = .05, we had a non-directional hypothesis, n = 10, and r = .331. After clicking calculate, G*Power tells you that the statistical power is .165293, which is extremely low. As in Chapter 18, the conventional level of statistical power is .80 or greater.

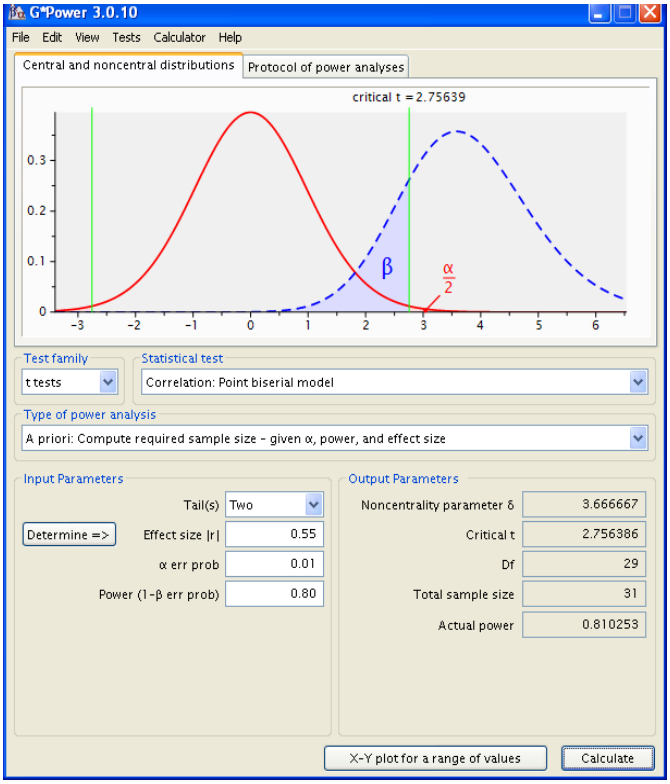
Recall from Section 15.7 that the significance test on the Ppearson correlation was non-significant. Because our power is so low we cannot conclude the relationship between age and books read per month this is truly non-significant. Stated differently, our non-significant correlation might be a Type II error. To determine whether this is a truly non-significant relationship or a Type II error we need to increase the sample size. If we increase the sample size and find the correlation is still non-significant, we can then conclude the relationship is non-existent.



We can also use G*Power in a priori power analyses of the Pearson correlation to determine the number of subjects needed to end up with a desired level of power. For example, assume am planning to collect data from a sample of subjects on two dependent variables. I know from prior studies the correlation between the variables is generally "moderate" and near *r* = .55. For my inferential test I am planning to use α = .01 and with a non-directional alternate hypothesis.

To determine the appropriate sample size, open G*Power 3 and under "Test family" be sure "t tests" is chosen, under "Statistical test" choose "Correlation: Point biserial method"; and under "Type of Power Analysis" choose "A priori: compute required sample size." Next, you'll need to enter the α-level, expected effect size, desired power, and choose "two" for Tail(s). For this example, assume I want Power = .80.

After clicking Calculate you are provided with the total sample size needed to achieve this amount of Power, which is *n* = 31 in this example. What if I wanted more Power, say, .95? Entering .95 into the Power areas, I am now told I need *n* = 45 subjects.



15.11 The Spearman Rank-Order Correlation (r_S)

In order to use the Pearson correlation, data must come from two quantitative variables measured on an interval or ratio scale. It may be the case that data from one or both variables is not interval or ratio, but from an ordinal scale. When data from one or both of your variables was measured on an ordinal scale, you must use the Spearman rank-order correlation (r_S) to measure the relationship between the variables.

The **Spearman rank-order correlation**, like the Pearson correlation, measures the degree to which two variables are related. In the case of the Spearman correlation the relationship reflects the degree to which the rank-ordering of the values for each variable agree or disagree. That is, because the data from one or both of the variables comes from an ordinal scale, which is the rank-ordering of entries for a variable, the Spearman correlation measures the degree to which the rank orderings for each variable are consistent. Generally, the Spearman correlation is used to assess the relationship between variables when:

- Both dependent variables are *quantitative* and at least one variable approximate an ordinal scale
- The two dependent variables are measured on the same individuals
- The observations on each variable are between-subjects in nature

Say we are college football fans and want to measure the relationship between order several teams finish at the end of a season (1^{st}, 2^{nd}, 3^{rd}) and the average number of points each team scored per game. The variable *Place of Finish* is ordinal, because a team finishes in 1^{st} place, 2^{nd} place, etc, but the variable Points per Game is ratio, because a team can score zero or more. This is okay because the Spearman correlation is used when at least one variable is on an ordinal scale; the other variable can be interval, ordinal, or ratio.

Team	Place	Points/Game
Springfield Isotopes	2	33
Shelbyville Shelbyvillians	5	25
Quahog Clams	3	30
Pawtucket Patriots	8	15
Evergreen Geoducks	1	35
Faber Mongols	4	31
PCU Whopping Cranes	7	17
Mongol-University Deathworms	6	27

It is okay to measure the relationship between an ordinal variable and a ratio/interval variable with the Spearman correlation, but you have to convert ratio or interval values into ordinal values. Look at the Points/Game variable and decide which value is largest, that value will be assigned an ordinal rank of 1. Next, decide which value is second largest, that value will be assigned an ordinal rank of 2. Etc. The table below shows what you have after converting the Points/Game values into rank-ordered values.

Team	Place	Points/Game (Ranked)
Springfield Isotopes	2	2
Shelbyville Shelbyvillians	5	6
Quahog Clams	3	4
Pawtucket Patriots	8	8
Evergreen Geoducks	1	1
Faber Mongols	4	3
PCU Whopping Cranes	7	7
Mongol-University Deathworms	6	5

The formula for the Spearman correlation is:

$$r_s = 1 - \frac{6(\sum D^2)}{N(N^2-1)}$$

In the formula above, N is the number of cases, or pairs, evaluated. In in this example, N = 8, because there are eight teams being evaluated. The value ΣD^2 is the *sum of the squared differences in rankings between the variables*. To compute this term you subtract the ordinal value for the variable on the right from the corresponding ordinal value from the variable on the left, then square those difference scores, and then add the squared differences. This is demonstrated in the D and D^2 columns in the table below:

Team	Place	Points/Game	D	D^2
Springfield Isotopes	2	2	0	0
Shelbyville Shelbyvillians	5	6	-1	1
Quahog Clams	3	4	-1	1
Pawtucket Patriots	8	8	0	0
Evergreen Geoducks	1	1	0	0
Faber Mongols	4	3	1	1
PCU Whopping Cranes	7	7	0	0
Mongol-University Deathworms	6	5	1	1
N = 8				$\sum D^2 = 4$

Once you have the D^2 value, you can insert that and the value of N into the formula above to calculate the Spearman correlation. This is done below:

$$r_s = \frac{6(4)}{8(8^2-1)} = 0.952$$

As was the case with the Pearson correlation, it is customary to report the Spearman rank-order correlation to two decimal places. Thus, the correlation between Place of Finish and point Per Game is $r_S = .95$. The larger the Spearman correlation, the greater is the consistency in the ranking (ordinal) values between the two variables. Importantly, positive values of the Spearman correlation indicate agreement in the rankings; that is, the rankings are highly similar for both variables. In contrast, a negative Spearman correlation indicates disagreement in the rankings; that is, the rankings are very different and are in opposite directions.

15.12 Statistical Significance of Spearman Correlation (r_S)

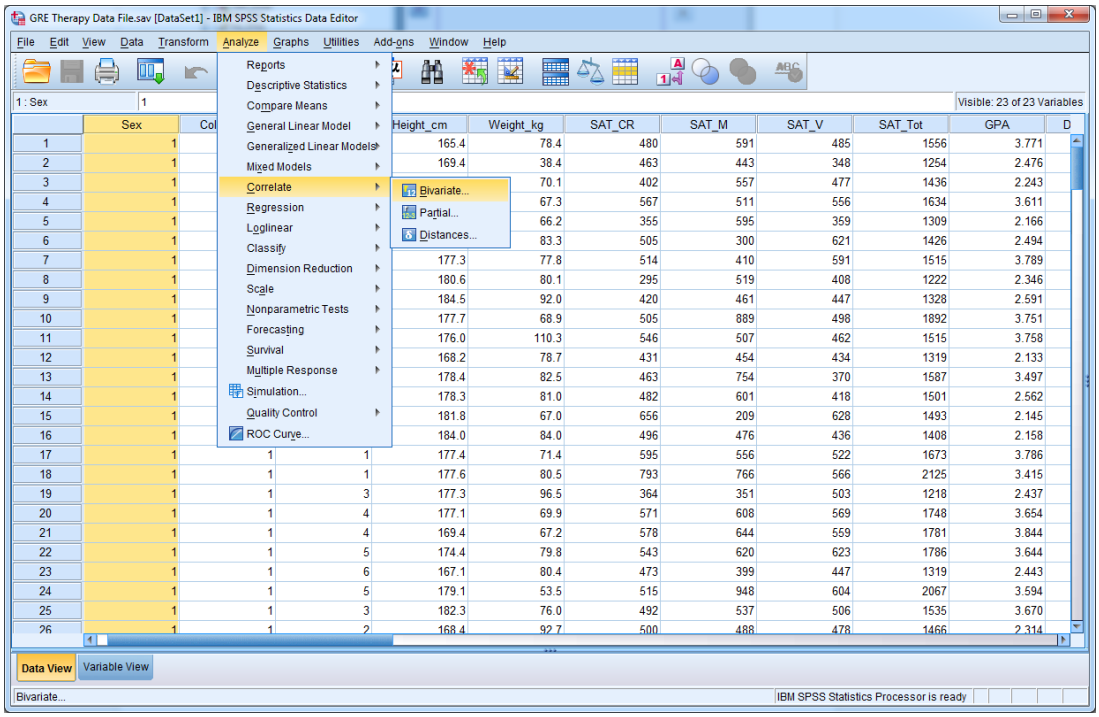
The statistical significance of the Spearman correlation can also be determined with a type of t-test:

$$t = r_s\sqrt{\frac{n-2}{1-r_s^2}} = 0.952\sqrt{\frac{8-2}{1-0.952^2}} = 7.606$$

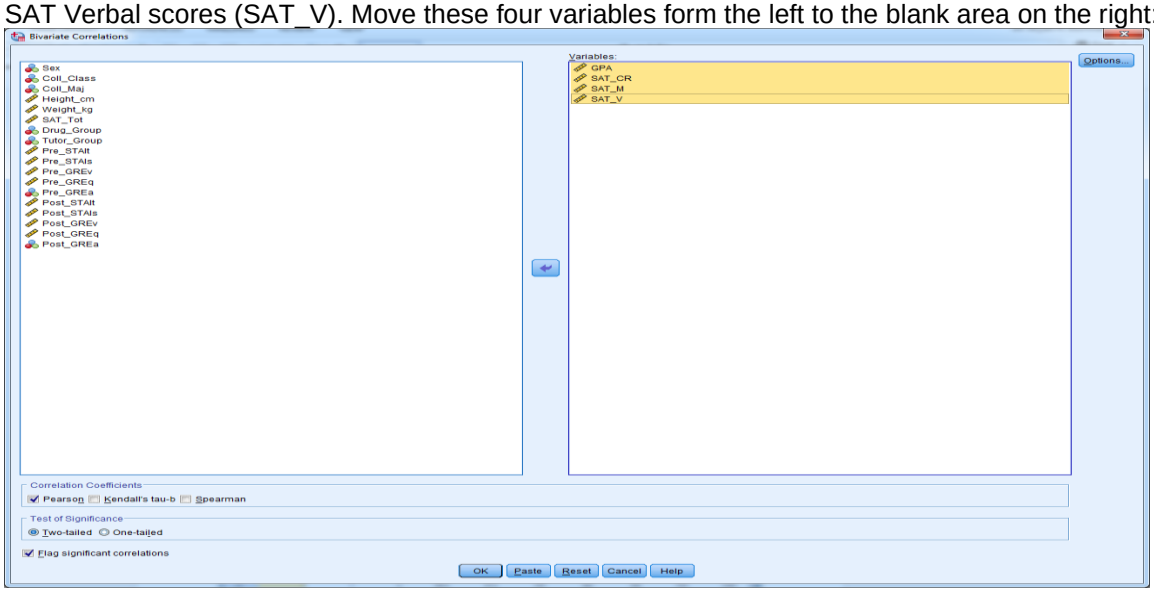
To determine the statistical significance of the Spearman correlation, we need to locate a *p*-value in Table t (t-tables), because this test is a variant of the t-test. Assume we are using a non-directional alternate hypothesis with an alpha level of α = .05. As was the case with the Pearson correlation, the degrees of freedom in the Spearmen correlation are df = n – 2, where n is the number of ordered pairs. Because the t-values in the t-tables do not go above 4.00, but the test statistic is equal to t = 7.607, we'll use t = 7.60. Looking up this test statistic value and df = 8, we find a p-value equal to $p < .0000$, which is less than the chosen alpha level (α = .05); hence, we conclude the Spearman correlation between place of finish and team points per game ($r_S = 0.952$) is statistically significant.

15.13 Pearson Correlation in SPSS

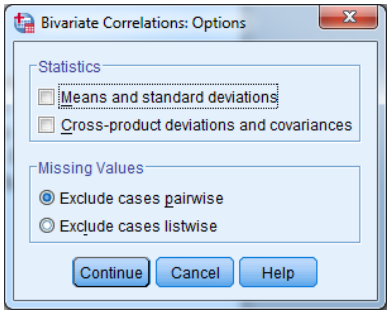
The following uses the GRE Therapy Data file. This data file is based on a hypothetical study examining the influences of a study-aid drug and types of tutoring on performance on the Graduate Record Examinations (GREs). To have SPSS perform correlations, from the Analyze menu, select Correlate, and then select Bivariate:



Remember, correlations are between two variables; however, with SPSS you can have the software compute correlations between multiple pairs of variables at the same time. For example, say we wanted to know the correlations between GPA, SAT Critical Reading scores (SAT_CR), SAT Math scores (SAT_M) and SAT Verbal scores (SAT_V). Move these four variables form the left to the blank area on the right:



SPSS will calculate the correlation between each pair of variables, and with four variables there are six correlations. Before clicking OK, click the Options button. If you want, you can have SPSS compute the descriptive statistics for each individual variable (Means and standard deviations), and calculate the sum of cross products and covariances between the variables (Cross-product deviations and covariances), but you should not worry about those now. In the main window, click the continue button to obtain the output, below:



Correlations

Correlations

		GPA	SAT_CR	SAT_M	SAT_V
GPA	Pearson Correlation	1	.531**	.436**	.532**
	Sig. (2-tailed)		.000	.000	.000
	N	240	240	240	240
SAT_CR	Pearson Correlation	.531**	1	-.062	.481**
	Sig. (2-tailed)	.000		.340	.000
	N	240	240	240	240
SAT_M	Pearson Correlation	.436**	-.062	1	-.048
	Sig. (2-tailed)	.000	.340		.461
	N	240	240	240	240
SAT_V	Pearson Correlation	.532**	.481**	-.048	1
	Sig. (2-tailed)	.000	.000	.461	
	N	240	240	240	240

**. Correlation is significant at the 0.01 level (2-tailed).

To locate a Pearson correlation in the table, cross reference the two variables you are interested in, by looking up one variable in the rows and the other variable in the columns. For example, to examine the correlation between GPA and SAT Math Scores, look up GPA the row headings and SAT_M in the column headings. You should find the Pearson correlation is r = .531, which is statistically significant (p < .001), and suggests as SAT Math scores increase, GPA also increases.

CH 15 Homework Questions

1. Draw a scatterplot for the following data:

Individual	X	Y
1	10	9
2	8	7
3	6	5
4	9	8
5	10	8
6	8	8
7	5	6

2. Draw a scatterplot for two variables that are negatively correlated.

3. Draw a scatterplot for two variables that are negatively correlated.

4. What does the magnitude of a correlation coefficient tell you about the relationship between two variables?

5. What does the sign of a correlation coefficient tell you about the relationship between two variables?

6. Under what conditions is Pearson correlations typically used to analyze a bivariate relationship?

7. What is the coefficient of determination? What does it tell you? What is the coefficient of alienation? What does it tell you?

8. State the critical t-Values for a test of the Pearson correlation under each of the following conditions:
   a. H_0: $\rho = 0$, H_1: $\rho \neq 0$, $\alpha = .05$, $n = 30$
   b. H_0: $\rho = 0$, H_1: $\rho > 0$, $\alpha = .05$, $n = 30$
   c. H_0: $\rho = 0$, H_1: $\rho \neq 0$, $\alpha = .01$, $n = 22$
   d. H_0: $\rho = 0$, H_1: $\rho < 0$, $\alpha = .01$, $n = 22$
   e. H_0: $\rho = 0$, H_1: $\rho \neq 0$, $\alpha = .05$, $n = 16$
   f. H_0: $\rho = 0$, H_1: $\rho \neq 0$, $\alpha = .01$, $n = 45$

9. Locate the p-values for a test of the Pearson correlation under each of the following conditions:
   a. H_0: $\rho = 0$, H_1: $\rho \neq 0$, $n = 20$, $t = 2.50$
   b. H_0: $\rho = 0$, H_1: $\rho > 0$, $n = 20$, $t = 2.50$
   c. H_0: $\rho = 0$, H_1: $\rho \neq 0$, $n = 36$, $t = 2.10$
   d. H_0: $\rho = 0$, H_1: $\rho < 0$, $n = 36$, $t = 2.10$

10. *Use the following data to answer the questions, below.*

Individual	X	Y
1	10	7
2	8	8
3	10	8
4	8	7
5	7	9
6	9	7
7	11	5
8	8	6
9	10	6
10	9	7

   a. Calculate the sum of squares for each variable and the sum of the cross products.
   b. Calculate the estimated standard deviations for each variable.
   c. Calculate the covariance.
   d. Calculate the Pearson correlation.
   e. Compute the coefficient of determination and the coefficient of alienation. What does the coefficient of determination tell you about the relationship?
   f. Calculate the t-value for the Pearson correlation.
   g. Assuming a non-directional hypothesis, what is the *p*-value for this correlation?
   h. Using an alpha level of $\alpha = .05$, is the Pearson correlation statistically significant?

i. Using G*Power, how much power was there to detect the relationship?

11. *Use the following data to answer the questions, below.*

Individual	X	Y
1	8	7
2	5	5
3	8	7
4	5	5
5	6	6
6	7	5
7	6	6
8	7	5
9	3	4
10	4	2
11	3	4
12	4	4

a. Calculate the sum of squares for each variable and the sum of the cross products.
b. Calculate the estimated standard deviations for each variable.
c. Calculate the covariance.
d. Compute the Pearson correlation.
e. Compute the coefficient of determination and the coefficient of alienation. What does the coefficient of determination tell you about the relationship?
f. Calculate the t-value for the Pearson correlation.
g. Assuming a non-directional hypothesis, what is the *p*-value for this correlation?
h. Using an alpha level of $\alpha = .05$, is the Pearson correlation statistically significant?
i. Using G*Power, how much power was there to detect the relationship?

12. *Use the following to answer the questions below.* Dr. Vader wants to measure the relationship between fear of the darkside of the Force and conformity. He randomly selects (i.e., threatens with Force Choke Hold) individuals from around the Galactic Empire. Each individual rates how much they fear the darkside of the force (X) on a scale of 1 – 10, where higher scores indicate more fear of the darkside, and each individual is measured on their willingness to conform to the Empire's demands (Y) on a scale of 1 - 10, where higher scores indicate more willingness to conform. The data from this sample appear below:

Individual	X	Y
Dan Solo	1	3
Flubacca	2	1
Duke Cloudwalker	7	9
Lando Griffin	5	6
Princess Vespa	7	9
Dark Helmet	1	2
President Scroob	6	8
Jabba is Nuts	5	7
Oil Can Notopen	2	1
Mr. Spot	4	5

a. In terms of ρ, state the hypotheses.
b. Calculate the mean of each variable.
c. Calculate the sum of squares for each variable and the sum of the cross products.
d. Calculate the estimated standard deviations for each variable.
e. Calculate the covariance.
f. Compute the Pearson correlation.

g. Compute the coefficient of determination and the coefficient of alienation. What does the coefficient of determination tell you about the relationship?
h. Calculate the t-value for the Pearson correlation.
i. Assuming a non-directional hypothesis, what is the *p*-value for this correlation?
j. Using an alpha level of α = .05, is the Pearson correlation statistically significant?

13. *Use the following to answer the questions below.* Recently, researchers examined the relationship between political attitude (liberal/conservative) and perceptual sensitivity (ability to detect small changes). One hundred sixty five students rated their political attitude on a scale of 1 – 11 and completed a perceptual sensitivity task. Sensitivity was measured by a metric called d' ("d-prime"), with larger values indicating greater sensitivity to detecting change. The researchers found a Pearson correlation between political attitude and d' of r = -.248.

a. Assuming the researchers made no prediction about the direction of the relationship, in terms of ρ, what are the null and alternate hypotheses?
b. Calculate the t-value for the Pearson correlation.
c. What is the *p*-value for this correlation?
d. Using an alpha level of α = .05, is the Pearson correlation statistically significant? What should these researchers conclude about the relationship?
e. Does the Pearson correlation represent a small, medium, or large relationship? Using G*Power, how much statistical Power did the researcher have for rejecting the null hypothesis?

14. *Use the following to answer the questions below.* Researchers examined the relationship between how much a person agrees with the Republican Party and authoritarian personality. They asked 177 students to rate how much they agreed with the Republican Party on a scale from 1 to 7 and gave them a test that assessed the authoritarian nature of each subject's personality, with larger scores indicating a more authoritarian personality. The researchers found a Pearson correlation between agreement with the Republican Party and the authoritarian personality test scores of r = .402. They assumed that the more one agrees with the Republican Party the higher their score would be on the authoritarian personality test

a. In terms of ρ, what are the null and alternate hypotheses?
b. Calculate the t-value for the Pearson correlation.
c. What is the *p*-value for this correlation?
d. Using an alpha level of α = .01, is the Pearson correlation statistically significant? What should these researchers conclude about the relationship?
e. Does the Pearson correlation represent a small, medium, or large relationship? Using G*Power, how much statistical Power did the researcher have for rejecting the null hypothesis?

15. *Use the following to answer the questions below.* Dr. Tomsurfrend, an educational psychologist, is interested in the relationship between poor time management and grades among high school seniors. He predicts that the more time a person spends on the social networking site Facebook, the lower their grades will be. He asks 50 high school seniors at Whitesboro Senior high school how much time they spend, per day, on the Facebook and also obtains the each student's GPA from the previous semester. The correlation is r = -.45 Using this information, answer the following questions:

a. In terms of ρ, what are the null and alternate hypotheses?
b. Calculate the t-value for the Pearson correlation.
c. What is the *p*-value for this correlation?
d. Using an alpha level of α = .01, is the Pearson correlation statistically significant? What should these researchers conclude about the relationship?

16. *Use the following to answer the questions below.* Dr. Hefner is trying to answer a long-debated question: 'Is shoe size statistically associated with penis length?' Dr. Hefner samples twenty males of the approximately same age, height, weight, and health. He measures each man's shoe size and, using a yardstick, measures each man's penis length (isn't it surprising that only twenty males signed up for this study?) The correlation between shoe size and penis length is +.08.

a. In terms of ρ, what are the null and alternate hypotheses?
b. Calculate the t-value for the Pearson correlation.
c. What is the *p*-value for this correlation?
d. Using an alpha level of α = .05, is the Pearson correlation statistically significant? What should these researchers conclude about the relationship?

17. *Use the following data to answer the questions below*. Below are the rankings of the top ten students at a school, as determined by the dean and the president of the school.

Student	Dean	President
Student A	10	10
Student B	4	6
Student C	2	8
Student D	3	3
Student E	9	7
Student F	8	9
Student G	1	4
Student H	6	2
Student I	5	1
Student J	7	5

a. Calculate the D and D^2 scores. What is the value of $\sum D$ and of $\sum D^2$?
b. Calculate the Spearman correlation.
c. Calculate the t-value for the Spearman correlation.
d. What is the *p*-value for this correlation?
e. Using an alpha level of α = .05, is the correlation statistically significant? What should the Dean and the President conclude about their rankings?

For Exercises 18 – 20, hypothetical values of *r*, *ρ*, and *n* are given. (a) Use Fisher's z-Transformation to determine the values of *r'* and *ρ'.* (b) Calculate $\sigma_{r'}$ (c) Calculate the test statistic (z-test) for the correlation. and (d) Determine the p-value. (e) Determine whether the value of *r'* is significant. Assume all hypotheses are non-directional (two-tailed) with α = .05.

18. *r* = -.25, *ρ* = -.10, *n* = 30

19. *r* = .86, *ρ* = .25, *n* = 20

20. c. *r* = .35, *ρ* = -.15, *n* = 10, α = .01

21. In a recent national poll, all registered voters were asked how much they agreed with the theory of evolution on a scale of 1 (strongly disagree) to 11 (strongly agree) and were also asked their overall political attitude on a scale of 1 (extremely liberal) to 11 (extremely conservative). The correlation was ρ = -.50, suggesting the more politically conservative someone is the less they tend to agree with the theory of evolution. Dr. Denis Leary is interested in whether this correlation is stable in various states, because some states are much more politically liberal than others and some are much more politically conservative than others. Dr. Leary samples 100 residents from Vermont who are registered voters and obtains their ratings to each of the two questions above. He finds that the correlation between agreement with the theory of evolution and political attitude is r = -.10 Use this information answer the following (assume α = .01 for all questions):

a. Dr. Leary wants to know whether the correlation obtained from his sample of Vermont voters is greater than the national correlation. In terms of ρ, state the hypotheses.
b. Determine the Fisher's Transformation of the r and ρ values.
c. Calculate $\sigma_{r'}$ for the sample of Vermont voters.
d. Calculate the z-Score of r'.

e. What is the p-value?
f. Is the difference between r and ρ statistically significant?

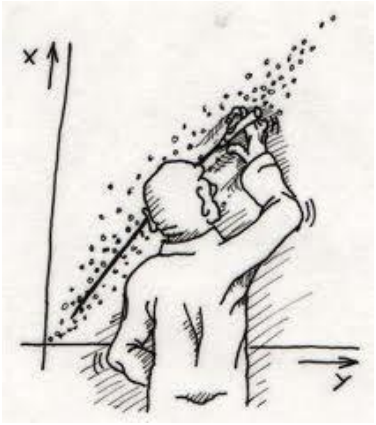
22. For each of the following situations, use G*Power to find the total number of subjects that would be needed to achieve the desired level of Power.

a. r = .50; α = .05 (non-directional); Power = .80
b. r = .10; α = .05 (directional); Power = .95
c. r = .80; α = .01 (directional); Power = .85
d. r = .25; α = .01 (non-directional); Power = .90
e. r = -.70; α = .01 (directional); Power = .85
f. r = -.25; α = .05 (directional); Power = .90

23. For each of the following situations, use G*Power, find the amount of achieved Power based on the parameters given.

a. r = .25; α = .05 (non-directional); *n* = 45
b. r = .55; α = .01 (non-directional); *n* = 110
c. r = -.35; α = .05 (directional); *n* = 30
d. r = -.45; α = .01 (directional); *n* = 25

Chapter 16: Regression



16.1 Prediction

Chapter 15 addressed how the relationship (correlation) between variables is measured. Once a correlation between variables has been established, it would be nice to find a way to predict values of one variable for given a value of another. Stated differently, if you find a relationship between variables, and if relationships indicate systematic changes in one variable as a function of another variable, you should be able to make predictions. For example, consider the data below. You should see that for every increase in X by 1 there is a corresponding increase in Y by 2, suggesting a positive linear relationship. Say we have a value of X = 8, which does not appear in the table. What value of Y would you predict is equal to X = 8? If the relationship between X and Y stays the same and every increase in X by 1 is associated with an increase in Y by 2, you should predict when X = 8 that Y = 22. How about X = 9? Again, if the relationship does not change, when X = 9 then Y = 24. How about when we know that Y = 8, what should X be equal to? Because we know the relationship between X and Y is positive and linear and for every change in Y by 2 there is a change in X by 1, when Y = 8 then X = 1.

X	Y
2	10
3	12
4	14
5	16
6	18
7	20

This chapter deals with **linear regression**, which is used to model the linear relationship between two variables, so values of one variable can be predicted from another variable. Specifically, given we have two variables in a relationship, linear regression attempts to fit a **linear model** to that relationship. If you look to the cartoon at the top of the page, you see a scatterplot depicting a positive linear relationship. The relationship is not perfectly linear and the individual in the cartoon is trying to plot a *best fit line* through the data. Linear regression attempts to fit a linear model (a straight line) to data, that is, to determine a mathematical expression of a relationship between variables.

Univariate regression uses one variable to model changes in another variable, and **multivariate regression** (**multiple regression**) uses several variables to model changes in another variable. In univariate regression there are two variables: a **dependent variable** (Y), which is also called the **regressed variable** or **predicted variable**, is the variable whose values we are predicting; and an **independent**

variable (X), which is also called the **regressor variable** or **predictor variable**, is the variable used to model changes in the dependent variable. In multivariate regression there is one dependent variable and two or more independent variables. Note, because relationships are never perfect, prediction will never be perfect. Regression procedures are designed so error in prediction is minimized through something called least-squares; a topic we will return to later.

16.2 The Regression Equation

In regression, values of one variable are used to predict values of a second variable. In this procedure you are creating a **regression equation** of a linear model, which takes this generic form:

$$\hat{Y} = b_0 + b_1 X$$

You should be familiar with this equation, because it is the equation for a straight line, which takes the mathematical form:

$$y = mx + b$$

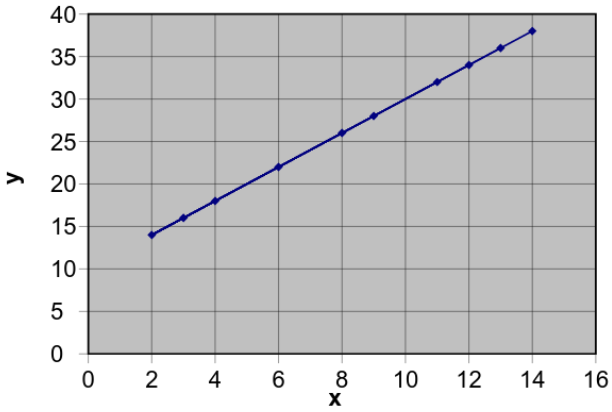
In the function, the **m** is a constant representing the change in y for every one-unit change in variable x; **b** is a constant that is expected when x = 0, and **x** is the variable used to predict y. Thus, the linear equation is modeling changes in y based on changes in x. For example, say we have the following linear equation, which is based on the data in the table below:

y = 2x + 10

The value 2 is the predicted change in y for every one-unit change in x, that is, if x increases by 1 then y is predicted to change by 2. The value 10, which was the constant b in the equation above, is the value of y when x = 0. Given this function we could plug in any value of x to calculate values of y. In the table to the right, for each given x, I have the corresponding y values calculated from the function above.

x	2x + 10	y
2	2(2) + 10	14
3	2(3) + 10	16
4	2(4) + 10	18
6	2(6) + 10	22
8	2(8) + 10	26
9	2(9) + 10	28
11	2(11) + 10	32
12	2(12) + 10	34
13	2(13) + 10	36
14	2(14) + 10	38

You can see that for every one-unit increase in x (from 2-3 or from 8-9), there is a corresponding two-unit increase in y (from 14-16 and 26-28, respectively). Thus, the value of m in the linear equation (b_1 in the regression equation) can be thought of as the **slope** of the linear equation. The slope of the linear equation can be better understood by plotting all of the x-y data points in a scatterplot and then drawing a line through them:



You can see that there is a straight line linking all of the x-y data points and that the slope (m = 2) shows a two-unit increase in y for every one-unit increase in x. Indeed, if you were to take any value of y from this graph (e.g., Y = 28) and call this y_1 and call its corresponding x value x_1 (X = 9), and then take any other y-value (Y = 26) and call this y_2 and its corresponding x value x_2 (X = 8), and substitute those values into the following equation you would find the outcome to be equal to the slope (m) of the linear equation:

$$m = \frac{y_1 - y_2}{x_1 - x_2} = \frac{28 - 26}{9 - 8} = 2$$

If you were to extend the straight line in the graph above so it passed through the ordinate, you would find it intersects at y = 10, which is the value of the constant b in the linear equation (b_0 in the regression equation). This value is the "y-intercept". The y-intercept is the value of y when x = 0. In a regression equation, Y' ('*y-prime*') is the predicted value of Y for a value of X; b_0 is the expected value of Y when X is equal to 0; b_1 is the predicted change in Y for every one-unit change in X (slope); and X is the independent variable that is used to model the changes in Y, the dependent variable. The following sections show how to calculate the slope (b_1) and the intercept (b_0) coefficients of the linear regression equation.

16.3 Calculating the Slope Coefficient (b_0)

The most important thing to keep in mind when calculating the slope and intercept in a regression equation is which variable you are regressing on the other, that is, which variable is being used to predict values of the other? The data from Chapter 15 are presented to the right where the correlation was calculated between age (X) and the number of books a person reads per month (Y). The descriptive statistics for each variable as are the summary statistics.

Name	Age (X)	Books per Month (Y)
A	22	8
B	20	4
C	22	4
D	21	3
E	35	6
F	32	7
G	38	4
H	40	6
I	20	3
J	40	5
n = 10	$\sum X$ = 290	$\sum Y$ = 50
	M_X = 29	M_Y = 5
SCP = 43	SS_X = 647	SS_Y = 26
cov = 4.778	$\hat{\sigma}^2_{\hat{\sigma}} = 71.889$	$\hat{\sigma}^2_{\hat{\sigma}} = 2.889$
r = 0.331	$\hat{\sigma}_{\hat{\sigma}} = 8.479$	$\hat{\sigma}_{\hat{\sigma}} = 1.7$

We want to fit a linear regression equation to these data to model the number of books a person reads per month based on age. Hence, we want to generate a regression equation to predict the number of books read per month from age.

Most terms needed to calculate the slope and intercept coefficients were calculated in Chapter 15, but this brings up an important point about linear regression: When calculating the slope and intercept values, you must calculate the slope first, because you need its value to calculate the intercept. Thus, start by calculating the slope regressing Y on X. The slope equation is:

$$b_1 = r(\frac{\widehat{s}_Y}{\widehat{s}_X})$$

In the formula, the estimated standard deviation of the dependent variable (Y) is divided by the estimated standard deviation of the independent variable (X), and the quotient is weighted by the correlation between the variables. Remember, the slope coefficient is the change in Y for every one unit change in X; thus, the dependent variable goes in the numerator and the independent variable goes in the denominator. Solving for the slope, we have:

$$b_1 = 0.331(\frac{1.7}{8.479}) = 0.066$$

Thus, Y (books read per month) is predicted to increase by 0.066 books for every one-unit increase in age (X). That is for every increase in a person's age by one year the number of books read per month is predicted to increase by 0.066 books. For example, if a person is 20 years old and that person reads 4 books per month, a 21 year old is predicted to read 4.066 books per month.

16.4 Calculating the Intercept Coefficient (b_0)

Once you have calculated the slope you can calculate the intercept (b_0). The equation is actually a variant of the regression equation that has been rewritten to solve for a:

$$b_0 = \overline{Y} - b_1\overline{X}$$

In this equation, you must determine the mean of Y, which is the mean of the regressed/predicted variable, and the mean of X, which is the mean of the independent variable. From the data in the table above, $\overline{X} = 29$ and $\overline{Y} = 5$. Plugging those values into the equation we get:

$$b_0 = 5 - (0.066)(29) = 3.086$$

This value (b_0 = 3.086) is the intercept in the regression equation, that is, the predicted value of Y when X = 0. Now that you have solved for the slope and intercept you can write out the regression equation by inserting the values of b_0 and b_1 into the generic regression equation from before:

$$Y' = 3.086 + 0.066X$$

You can use this equation to solve for predicted values of Y by plugging in any possible value of X and solving for Y', which we will do in the following section.

$$Y' = 3.086 + 0.066X$$

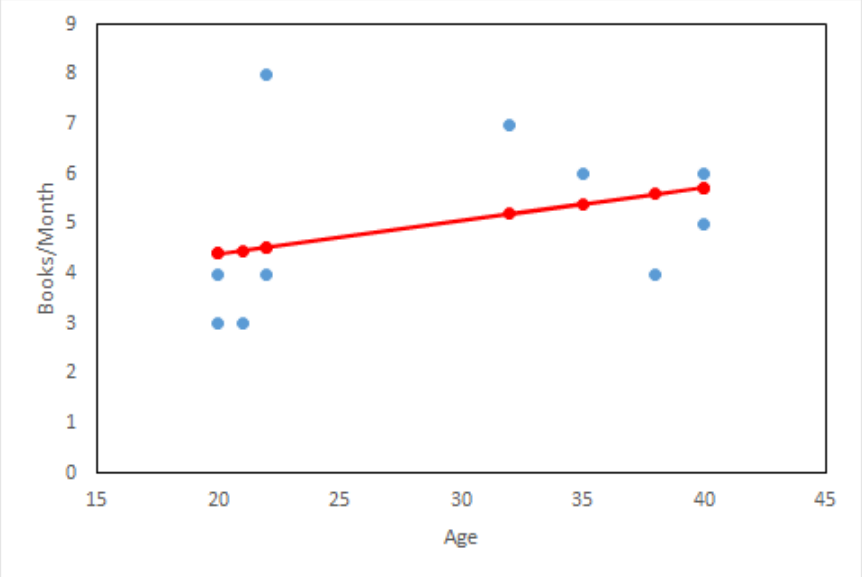
16.5 Plotting a Regression Line

The regression equation is the model of a straight line through the data points for two variables. Hence, we can plot the function (regression equation) in a scatterplot with the data. There are two ways you can do this: (1) Select a small value of X to compute a predicted value of Y and select a large value of X to predict a second value of Y. Plot both X-Y' data points in the scatterplot and join them by a straight line. (2) Compute each Y' value corresponding to each known X-value, plot each X-Y' data point in the scatterplot and draw a line through all of them.

In the table below, I calculated the predicted value (Y') for each X-value. Notice there are differences between the actual y-values and the predicted y'-values. This is because the regression equation is predicting values of Y. Because the relationship between X and Y is not perfectly linear (as indicated by r = 0.331), prediction will not be perfect:

Name	Age (X)	Books (Y)	Y' = 3.086 + 0.066X	Y'
A	22	8	3.086 + .066(22)	4.538
B	20	4	3.086 + .066(20)	4.406
C	22	4	3.086 + .066(22)	4.538
D	21	3	3.086 + .066(21)	4.472
E	35	6	3.086 + .066(35)	5.396
F	32	7	3.086 + .066(32)	5.198
G	38	4	3.086 + .066(38)	5.594
H	40	6	3.086 + .066(40)	5.726
I	20	3	3.086 + .066(20)	4.406
J	40	5	3.086 + .066(40)	5.726

Plotting the X - Y' data points in a scatterplot with the actual X-Y data points, we get the graph below. In the scatterplot, the actual X-Y data points are plotted in blue and the X-Y' data points are plotted in red with a line passing through them. This is the 'best fit' linear relationship between age (X) and number of books read per month (Y). Again, notice that the Y' values differ from the actual Y values. Again, with an imperfect linear relationship, this is to be expected. The differences between the Y values and predicted Y' values is residual variance, which we will turn to in the next section.



16.6 Residual Variance and Standard Error of the Estimate

Before getting into how residual variance is calculated let me explain a little about what residual variance is and where it comes from. The total variability in the scores of the predicted variable (Y) is simply the sum of squares for that variable (SS_Y). This is composed of two sources of variance: variability due to random sampling error and variability due to regression. Specifically:

Total Variability = Regression Variability + Error Variability

If total variability in the predicted variability is the sum of squares for that variable (SS_Y), then the total error variability must be a sum of squares for error, which we call **sum of squares residual** (**SS_{Resid}**). Also, the regression variability must be a sum of squares due to regression, which we call **sum of squares regression** (**SS_{Reg}**). Thus, total variability is equal to sum of squares residual added to the sum of squares regression:

$$SS_Y = SS_{Reg} + SS_{Resid}$$

Computationally, this is:

$$\sum(Y-\overline{Y})^2 = \sum(Y'-\overline{Y})^2 + \sum(Y-Y')^2$$

Sum of squares-regression (SS_{Reg}) is the total variance in the predicted scores (Y') that can be explained in the linear relationship between X and Y. You can see this in the first term after the equal sign in the equation above: The sum of squares regression is the total variability between the predicted values (Y') and the mean of the actual y-values ($\overline{Y}$). Hence, sum of squares residual is measuring how close the predicted scores (Y') are to the actual mean. The sum of squares residual (SS_{Resid}) is the total variance in Y that cannot be explained in the linear relationship between X and Y; thus, it is unexplained variance. You can see this in the second term after the equal sign in the equation above: The sum of squares residual is the variability between the predicted y-values (Y') and the actual y-values (Y). Hence, sum of squares residual is measuring how closely the predicted scores (Y') are to the actual scores (Y), where any deviation must be due to random, unexplainable error.

The components to the regression equation (b_0 and b_1) are calculated in such a way that residual variance is minimized in what is called the **least-squares criterion**. What we want to know is how much error there is in predicting Y from X; that is, how much unexplained variation there is in the predicted values of Y.

To calculate sum of squares regression, use the following formula:

$$SS_{Reg} = r^2SS_Y$$

To calculate sum of squares residual, use the following formula:

$$SS_{Resid} = (1 - r^2)SS_Y$$

From data in the preceding sections, the sum of squares for the number of books a person read per month was $SS_Y = 26$ and $r = 0.331$. Plugging these values into the equations above, we have:

$$SS_{Reg} = r^2SS_Y = (0.331^2)26 = 2.86$$

$$SS_{Resid} = (1 - r^2)SS_Y = (1 - 0.331^2)26 = 23.14$$

These values are summed, or total variance values. What we want is the average residual variance. The formula for **average residual variance** is found by taking the average of the sum of squares residual:

$$\widehat{s}^2_{Y'} = \frac{\sum(Y-Y')^2}{n-2} = \frac{(1-r^2)}{n-2} = \frac{SS_{Resid}}{n-2}$$

You can see that residual variance is the sum of squares residual divided by degrees of freedom, which is n – 2 in this case, because there are two dependent variables. Note that each of these formulas are the same way of symbolically writing 'the average sum of squares residual'; that is, the numerator in each case is just the sum of square residual. Solving for the average residual variance, we have:

$$\widehat{s}^2_{Y'} = \frac{23.14}{n-2} = 2.892$$

This value is the variation in Y that cannot be explained in the linear relationship. If we wanted to know the average deviation between each actual Y value and the predicted value of Y we would take the square-root of residual variance, which will give us the **standard error of the estimate**. That is, the standard error of the estimate is simply the square root of the average residual variance that was just calculated:

$$\widehat{s}_{Y'} = \sqrt{\frac{\sum(Y-Y')^2}{n-2}} = \sqrt{\frac{(1-r^2)}{n-2}} = \sqrt{\frac{SS_{Resid}}{n-2}} = \sqrt{\widehat{s}^2_{Y'}}$$

The term under the radical sign is just the formula for average residual variance. Conceptually, the standard error of the estimate is the square root of residual variance. The standard error of the estimate is functionally equivalent to a standard deviation. Remember, the standard deviation is the average difference between a score and the mean of a distribution; thus, the standard error of the estimate is the average deviation between the actual y values and the predicted values:

$$\widehat{s}_{Y'} = \sqrt{\widehat{s}^2_{Y'}} = \sqrt{2.892} = 1.7$$

Thus, for this set of data, the predicted values (Y') are estimated to deviate from the actual values (Y) by 1.7 units (books read per month).

16.7 Homoscedascity

In order for residual variance and the standard error of the estimate to accurately measure the error in prediction, we must make an assumption of homoscedascity. Homoscedascity is seen when the actual Y-values vary to the same degree at each level of X; that is, the variance in Y for every value of X is equivalent. Heteroscedascity occurs when the variance in Y is not equal at every level of X. Homoscedascity assumes that the values of Y are normally distributed at each value of X; hence, there is equal variance in Y at each value of X. If the data is heteroscedastic, it means the estimate of the residual variance and the standard error of the estimate does not accurately measure the variability between the actual Y-value and the predicted values.

16.8 Making Inferences about Regression Equations and Coefficients

There are tests of statistical significance for regression coefficients (intercept and slope) and for the entire regression equation. There are several things that can be significant in any regression equation: First, the entire regression equation can be a significant model of a linear relationship. Second, the slope (b_1) can be a significant predictor for changes in Y. Finally, the intercept (b_0) can be significant (which is usually assumed to be zero under the null). Instead of burrowing into the various assumptions and population

distributions about the slope, intercept and regression equation, I'll focus more on how significance is determined for the regression equation and components that we calculated in earlier sections.

Recall that the regression equation that predicted number of books read per month (Y) from age (X) was:

$$Y' = 3.086 + 0.066X$$

We will evaluate the statistical significance of the entire regression equation, the slope, and the intercept. We'll start with the entire regression equation, because some of that information will be necessary for evaluating the significance of the slope and intercept coefficients.

We can conclude the regression equation is a significant predictor of the number of books read per month from age if a significant proportion of the variance in Y (books read per month) can be explained by variation in X (age). Recall, that the total variation in the predicted variable (SS_Y), which we'll call the **sum of squares total** (**SS_{Total}**) is a combination of the total variance that can be explained in the relationship between Y and X (SS_{Reg}) and the total error, or residual variance, that cannot be explained (SS_{Resid}).

Recall that the sum of squares regression (SS_{Reg}) and is calculated by summing the squared deviations between the predicted values (Y') and the mean of Y:

$$SS_{Reg} = \sum (Y' - \overline{Y})^2 = r^2 SS_Y$$

The sum of squares residual (SS_{Resid}) and is calculated by summing the squared deviations between the actual values (Y) and the predicted values (Y'):

$$SS_{Resid} = \sum (Y - Y')^2 = (1 - r^2) SS_Y$$

Total variance is the sum of squares total (SS_{Total}) and is calculated by summing the squared deviations between the actual values (Y) and the mean of the Y values. Thus:

$$SS_{Total} = \sum (Y - \overline{Y})^2$$ □□□□□□ = ∑ $(□ - \underline{□})^2$

The total variance in the predicted variable is equal to (can be partitioned into) total residual variation and the total variation due to regression:

$$\sum (Y - \overline{Y})^2 = \sum (Y' - \overline{Y})^2 + \sum (Y - Y')^2$$

Or

$$SS_Y = SS_{Reg} + SS_{Resid}$$

These three sources of variance were obtained in earlier sections and chapters: $SS_{Total} = SS_Y = 26$; $SS_{Reg} = 2.86$; and $SS_{Resid} = 23.14$.

Next, we analyze whether the variance accounted for in the relationship between number of books red per month and age (SS_{Reg}) is significantly greater than the variance that cannot be accounted for in this relationship (SS_{Resid}). Thus, we are going to analyze variance in this relationship. Recall that all inferential statistics boil down to a ratio of some amount effect (or explained variance) to some amount of random error (unexplained variance). Thus, we can determine the significance of the regression equation by forming a ratio of the variance due to regression over the residual variance:

$$\frac{\sigma^2_{Regression}}{\sigma^2_{Residual}}$$

However, we cannot place the sums of squares regression and residual into this test statistic, because we need to take into account the degrees of freedom associated with each source of variance. That is, we need to divide first each sum of squares above by their associated degrees of freedom to get the 'mean variance' of each source of variance. This 'mean variance' is called a **mean square** (**MS**).

The degrees of freedom associated with the regression variance (df_{Reg}) are always equal to k – 1, where k is the number of coefficients in the regression equation. In our example, there are k = 2 coefficients; the slope (b_1) and the intercept (b_0). Hence, in this example, df_{Reg} = 2 – 1 = 1. This degrees of freedom is also known as the degrees of freedom due to the effect of a variable (df_{Effect}) and also the degrees of freedom between groups (df_B). The degrees of freedom associated with the residual variance (df_{Resid}) are equal to n – 2; where n is the number of subjects in the data set. In this example, there are n = 10 subjects, so the degrees of freedom associated with residual variance are df_{Resid} = 10 – 2 = 8. This degrees of freedom is also known as the degrees of freedom due to error (df_{Errort}) and the degrees of freedom within groups (df_W).The total degrees of freedom (df_{Total}) are equal to n – 1; thus, 10 – 1 = 9. Note that df_{Total} = df_{Reg} + df_{Resid} (9 = 1 + 8). The degrees of freedom regression and the degrees of freedom residual are used to determine the statistical significance of the regression equation.

Now we can calculate the mean squares. To simplify, a mean square is just a sum of squares (total variability) divided by its associated degrees of freedom; thus:

$$MS = \frac{SS}{df}$$

The **mean square for regression** is the sum of squares regression divided by the degrees of freedom for regression and the **mean square for residual variance** is the sum of squares residual divided by the degrees of freedom residual. We only need to calculate the mean square for the regression variance and residual variance, which I have done below:

$$MS_{reg} = \frac{SS_{Reg}}{df_{Reg}} = \frac{2.86}{1} = 2.86 \qquad MS_{Resid} = \frac{SS_{Resid}}{df_{Resid}} = \frac{23.14}{8} = 2.892$$

Note, the mean square residual is also the average residual variance ($\hat{\sigma}^2_{\varepsilon'}$) from Section 15.6; hence, the square root of the mean square residual will give you the standard error of the estimate ($\hat{\sigma}_{\varepsilon'}$). The test statistic for determining significance of the regression equation is the regression (explained) variance over the residual (error) variance. This test statistic is called the F-Ratio:

$$F = \frac{MS_{Reg}}{MS_{Resid}} = \frac{2.86}{2.891} = 0.989$$

This is the test statistic that will be used to determine the statistical significance of the regression equation. To determine significance we need to look-up a *p*-value in Table 3 in Appendix A, which are the probabilities associated with areas under the F-Distribution (more on this in the following chapters). There are actually six tables associated Table 3, Table 3-A through Table 3-F, with each table being associated with a different number of degrees of freedom due to regression (df_{Reg}). In this example, we have df_{Reg} = 1, so we'll use Table 3-A, which is associated with df_B = 1. A portion of Table 3-A is reproduced, below.

Table 3-A: df_B = 1

	Degrees of Freedom Error (df_W)																								
F	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**14**	**16**	**18**	**20**	**22**	**24**	**26**	**28**	**30**	**40**	**45**	**50**	**60**	**80**	**100**	**250**	**1000**
1.00	.3739	.3632	.3559	.3506	.3466	.3434	.3409	.3370	.3343	.3322	.3306	.3293	.3282	.3273	.3265	.3259	.3253	.3233	.3227	.3221	.3213	.3203	.3197	.3183	.3176
1.10	.3535	.3423	.3347	.3291	.3249	.3216	.3190	.3149	.3120	.3099	.3081	.3068	.3057	.3047	.3039	.3032	.3026	.3006	.2999	.2993	.2985	.2974	.2968	.2953	.2945
1.20	.3349	.3233	.3153	.3096	.3052	.3018	.2990	.2948	.2918	.2895	.2878	.2863	.2852	.2842	.2834	.2827	.2820	.2799	.2792	.2786	.2777	.2766	.2760	.2744	.2736
1.30	.3178	.3059	.2977	.2917	.2872	.2836	.2808	.2765	.2733	.2710	.2692	.2677	.2665	.2655	.2646	.2639	.2632	.2610	.2602	.2596	.2587	.2576	.2569	.2553	.2545
1.40	.3022	.2899	.2815	.2753	.2707	.2670	.2641	.2596	.2564	.2540	.2521	.2506	.2494	.2483	.2474	.2467	.2460	.2437	.2429	.2423	.2414	.2402	.2395	.2378	.2370
1.50	.2879	.2752	.2666	.2603	.2555	.2518	.2487	.2442	.2409	.2384	.2365	.2349	.2336	.2326	.2317	.2309	.2302	.2278	.2270	.2264	.2255	.2243	.2235	.2218	.2210
1.60	.2746	.2617	.2528	.2464	.2415	.2377	.2346	.2299	.2266	.2240	.2220	.2204	.2191	.2180	.2171	.2163	.2156	.2132	.2124	.2118	.2108	.2096	.2088	.2071	.2062

1.70	.2623	.2491	.2401	.2335	.2286	.2246	.2215	.2167	.2133	.2107	.2087	.2071	.2058	.2047	.2037	.2029	.2022	.1997	.1989	.1983	.1973	.1960	.1953	.1935	.1926
1.80	.2508	.2374	.2283	.2216	.2165	.2126	.2094	.2046	.2011	.1984	.1964	.1947	.1934	.1923	.1913	.1905	.1898	.1873	.1864	.1858	.1848	.1835	.1828	.1809	.1800
1.90	.2402	.2266	.2173	.2105	.2054	.2014	.1981	.1932	.1897	.1870	.1850	.1833	.1819	.1808	.1798	.1790	.1783	.1757	.1749	.1742	.1732	.1719	.1712	.1693	.1684
2.00	.2302	.2164	.2070	.2002	.1950	.1909	.1877	.1827	.1792	.1765	.1744	.1727	.1713	.1701	.1692	.1683	.1676	.1650	.1642	.1635	.1625	.1612	.1604	.1585	.1576

Assume we select an alpha level of α = .05. First, note that there is no such thing as a directional or non-directional hypothesis in the F-Distribution; hence, once we find the p-value associated with the test statistic (0.989) we always compare it directly to the selected alpha level. To determine the statistical significance of the regression equation lookup the value of the *F*-Ratio (0.989, which rounds to 1.00 in Table 3-A) in the leftmost column (highlighted in yellow) and lookup the value for the degrees of freedom residual (df_{Resid} = 8) in the column headings (highlighted in yellow). The p-value associated with this F-Ratio is 0.3466. This value is greater than the selected alpha level (.05), so we conclude the regression equation is a not significant predictor of books red per month. That is, this linear regression equation based on age does not account for a significant amount of variance in the number of books read per month based on age.

Although the regression equation was been found to be non-significant, you could evaluate whether the slope (b_1) and intercept (b_0) are statistically significant. Note that the interpretations of the significance of these coefficients can be misleading if the regression equation is not significant, as was the case here. Nonetheless, below, I show how to determine statistical significance of these regression coefficients. For both the slope and the intercept, we will conduct a t-test. Each t-test is the difference between the calculated regression coefficient and its expected value under the null hypothesis, over a standard error. The expected value of each regression component under the null hypothesis, denoted β_1 for the slope and β_0 for the intercept, are usually expected to be zero, but they do not have to be. The standard error for the slope, $SE(b_1)$, is:

$$SE(b_1) = \frac{\widehat{s}_{Y'}}{\sqrt{SS_X}} = \frac{1.7}{\sqrt{647}} = 0.067$$

The standard error for the intercept, $SE(b_0)$, is:

$$SE(b_0) = \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{SS_X}} = \sqrt{\frac{1}{10} + \frac{29^2}{647}} = 0.792$$

Both t-tests take the same general form:

$$t = \frac{b_i}{SE(b_i)}$$

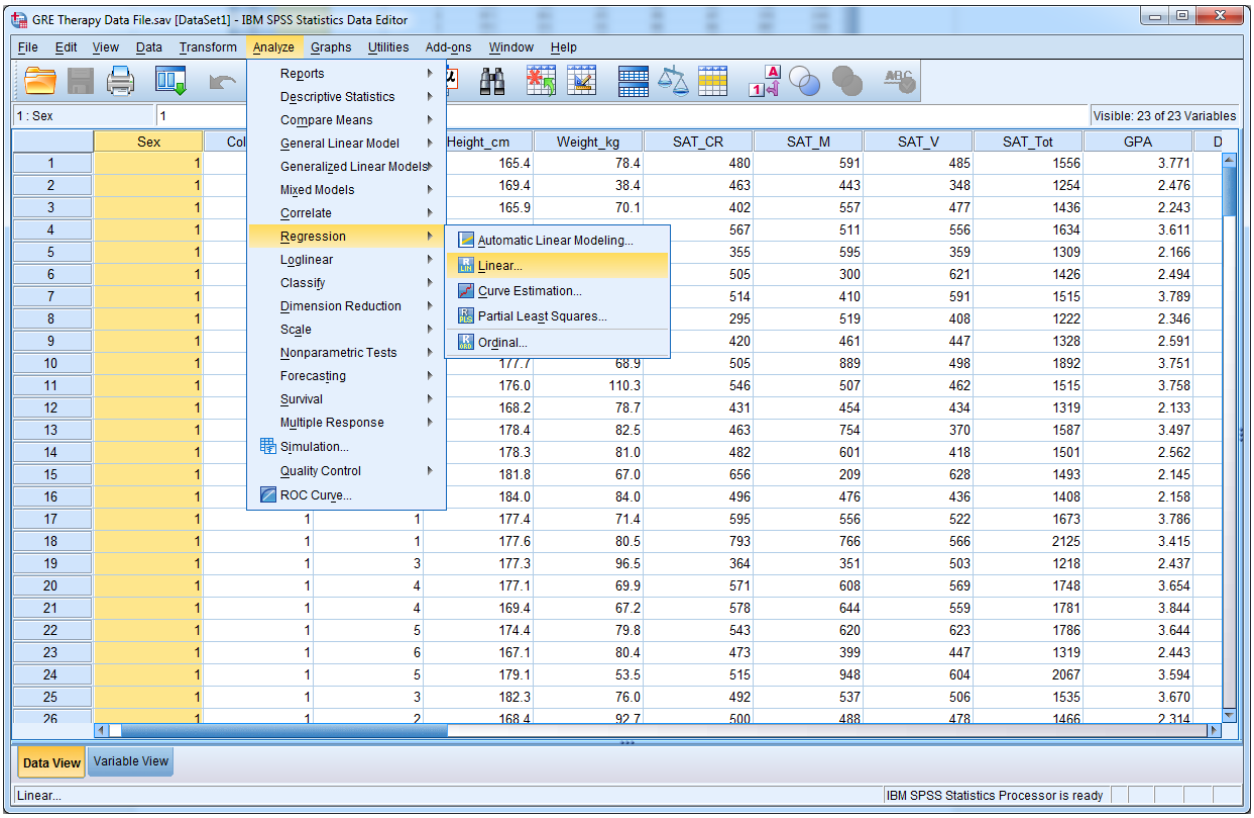
Plugging in the coefficients and the standard errors into the t-test formulas, we get (the slope and intercept coefficient values come from the regression equation earlier):

$$t = \frac{b_1}{SE(b_1)} = \frac{0.066}{0.067} = 0.985 \quad \text{and} \quad t = \frac{b_0}{SE(b_0)} = \frac{3.086}{0.792} = 3.896$$

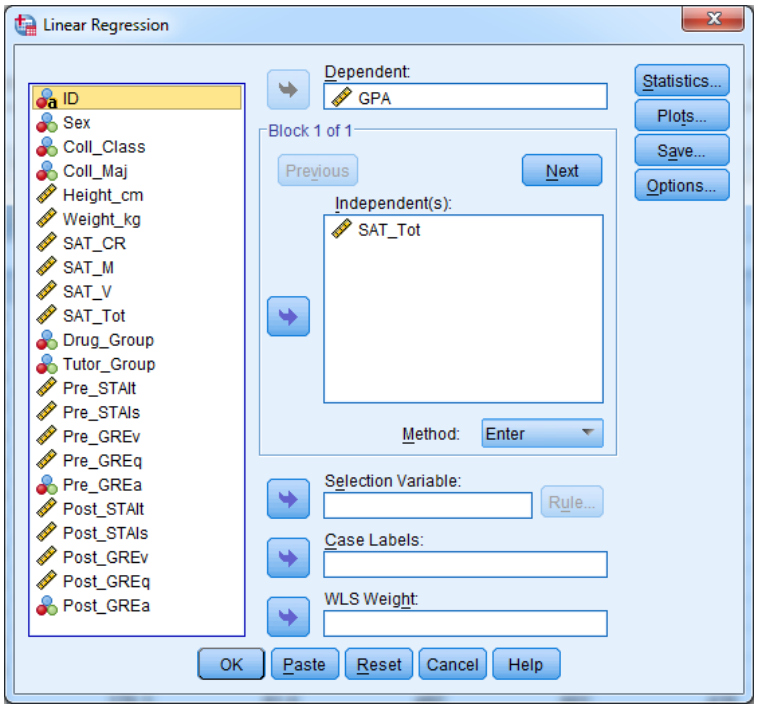
To evaluate the statistical significance of each regression coefficient, look up the t-values in Table 2 using df = 8. For the slope (t = 0.985, using t = 0.80 in Table 2), the two-tailed p-value is *p* = .4468, indicating the slope is not statistically significant. This means is that there is not a significant increase in the number of books red per month based on an increase in a person's age. For the intercept (t = 3.896, using t = 3.90 in Table 2) the p-value is *p* = .0046, which is less than our alpha level (.05), indicating the intercept if statistically significant. This means the intercept is significantly different from zero and that this sample reads a significant number of books per month, compared to not reading any books per month.

16.9 Regression in SPSS

The following uses the GRE Therapy Data file. This data file is based on a hypothetical study examining the influences of a study-aid drug and types of tutoring on performance on the Graduate Record Examinations (GREs). To have SPSS perform regression, from the Analyze menu, select Regression, and then select Linear:



In the window that opens, there is an area to declare the dependent variable and an area to declare one or more independent variables. In this example, assume we treat the GPA as the dependent variable and SAT_Tot as the independent variable. If you click the Statistics button, you can ask for descriptive statistics to be calculated as well. When you have done so, click the OK button.



The SPSS output should resemble the output below. Description of the output follows the SPSS output.

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
GPA	3.01082	.670476	240
SAT_Tot	1504.62	190.169	240

Correlations

		GPA	SAT_Tot
Pearson Correlation	GPA	1.000	.774
	SAT_Tot	.774	1.000
Sig. (1-tailed)	GPA	.	.000
	SAT_Tot	.000	.
N	GPA	240	240
	SAT_Tot	240	240

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	SAT_Tot[b]	.	Enter

a. Dependent Variable: GPA

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.774[a]	.599	.597	.425693

a. Predictors: (Constant), SAT_Tot

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	64.311	1	64.311	354.887	.000[b]
	Residual	43.129	238	.181		
	Total	107.440	239			

a. Dependent Variable: GPA

b. Predictors: (Constant), SAT_Tot

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.093	.220		-4.979	.000
	SAT_Tot	.003	.000	.774	18.838	.000

a. Dependent Variable: GPA

The first table in the output (Descriptive Statistics) provides the means, standard deviations, and sample sizes for each variable in the regression model. The second table (Correlations) lists the Pearson correlations between each pair of variables in the regression model. In this case, because there are only two variables in the model (the independent and dependent variables), there is only one Pearson correlation. The third table (Variables Entered/Removed) indicates which variables are the independent variables and which is the dependent variable in the model. The fourth table (Model Summary) provides the correlation between all variables in the model (R), which is just the Pearson correlation in this case, R^2, which is the coefficient of determination, and the standard error of the estimate. The fifth table (ANOVA) provides the sums of squares, degrees of freedom, mean squares, F-Ratio (F) and p-value (Sig.) for the analysis of variance on the entire regression equation. Finally, the sixth table (Coefficients) provides slope and intercept coefficients (under B), their standard errors, and the results of the t-tests on the coefficients.

In this example, the slope coefficient is $b_1 = 0.003$ and the intercept coefficient is $b_0 = -1.093$. Hence, the regression equation is $Y = -1.093 + 0.003X$.

CH 16 Homework Questions

1. What is the linear regression line (model)?

2. Describe what information is conveyed by the slope coefficient.

3. Describe what information is conveyed by the intercept coefficient.

4. What is the generic form of the linear regression model?

5. Assume you have a perfect linear relationship between variables X and Y and the relationship has a slope of 2.5. By how many units does Y change if X changes by 1 Unit? If X changes by 3 units? If X changes by 8 units?

6. What information is provided by the standard error of the estimate? What measure of variability is the standard error of the estimate similar to?

7. *Use the following to answer the questions below.* A researcher collected data on variables X any Y from each of ten subjects. The data for each of these ten subjects are below, as are the summary statistics and Pearson correlation between the two variables:

i	X	Y
A	10	0
B	9	2
C	7	1
D	6	2
E	6	2
F	4	4
G	3	6
H	3	8
I	2	7
J	0	8

$\overline{X} = 5$ $SS_X = 90$ $\hat{s}_X^2 = 10$ $\hat{s}_X = 3.162$

$\overline{Y} = 4$ $SS_Y = 82$ $\hat{s}_Y^2 = 9.111$ $\hat{s}_Y = 3.018$

SCP = -79 $cov_{xy} = -8.778$ $r = -0.920$

a. Calculate the slope coefficient of a linear equation regressing Y on X (predicts Y from X).
b. Calculate the intercept coefficient for a linear equation regressing Y on X.
c. State the regression equation based on the coefficients just calculated.
d. Calculate the predicted values of Y (Y') for each of the following individuals.
   Individual B
   Individual H
   Individual E
e. Calculate the sum of squares residual and the sum of squares regression (using the computational methods).
f. Calculate the standard error of the estimate. What does this value tell you about the predicted values of Y?

8. *Use the following to answer the questions below:* A researcher collected data on variables X any Y from each of ten subjects. The data for each of these ten subjects are below, as are the summary statistics and Pearson correlation between the two variables:

Individual	X	Y
A	3	7
B	8	9
C	3	3
D	2	8
E	6	8
F	6	9
G	8	6
H	5	4
I	7	2
J	2	4

$\overline{X} = 5$ $SS_X = 50$ $\hat{s}_X^2 = 5.556$ $\hat{s}_X = 2.357$

$\overline{Y} = 6$ $SS_Y = 60$ $\hat{s}_Y^2 = 6.667$ $\hat{s}_Y = 2.582$

SCP = 10 $cov_{xy} = 10$ $r = 0.182$

a. Calculate the slope coefficient of a linear equation regressing Y on X (predicts Y from X).
b. Calculate the intercept coefficient for a linear equation regressing Y on X.
c. State the regression equation based on the coefficients just calculated.
d. Calculate the predicted values of Y (Y') for each of the following individuals.
   Individual A
   Individual G
   Individual J
e. Calculate the sum of squares residual and the sum of squares regression (using the computational methods).
f. Calculate the standard error of the estimate. What does this value tell you about the predicted values of Y?

9. Based on the standard errors of the estimated calculated in Exercises 7 and 8, which regression equation is a better predictor of variable Y?

10. *Use the following to answer the questions below*: A professor has noticed that students who lack basic math skills do not perform well in his statistics course. The professor creates a math pretest that has a range of zero to ten. The test is given at the first class meeting to determine whether each student has the math skills necessary for the statistics course (n = 8). To evaluate the validity of his pretest, at the end of the semester the professor regresses student course grade points (Y) on the pretest scores (X). The regression equation and statistics appear in the table below.

Pretest (X)	Grade Points(Y)	Y' = 0.884 + 0.326X
$\overline{X} = 5$	$\overline{Y} = 2.675$	$r = 0.744$
$SS_X = 82$	$SS_Y = 17.715$	cov = 26.7

a. Calculate the sum of squares regression and the sum of squares residual.
b. Determine the degrees of freedom regression, degrees of freedom residual, and degrees of freedom total.
c. Calculate the mean square residual and the mean square regression

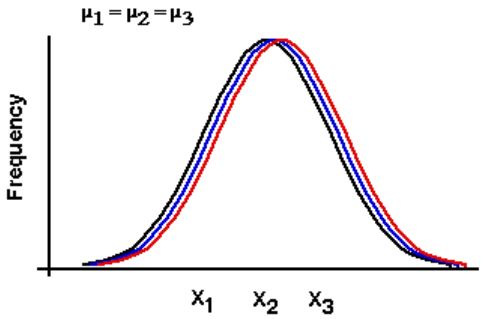
d. Calculate the F-Ratio.
e. Determine the *p*-value for the F-Ratio.
f. Using α = .05, is the regression equation statistically significant? Why?
g. Calculate the standard error of the estimate from the average residual variance ($MS_{Residual}$).
h. Calculate the standard error of the slope $SE(b_1)$ and standard error of the intercept $SE(b_0)$.
i. Calculate the *t*-Value for the slope (b_1). What is the p-value?
j. Calculate the *t*-Value for the intercept (b_0). What is the p-value?

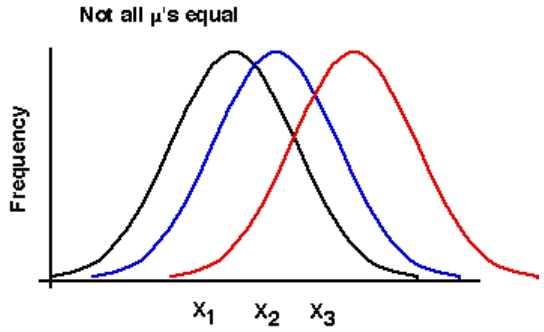
11. *Use the following to answer the questions below*: Using the data from Chapter 15 Exercise 12, Dr. Vader wants to predict a person's willingness to conform (Y) from their self-reported fear of the darkside of the force (X). Recall, that each of n = 10 individuals rated how much they feared the darkside of the force (X) on a scale of 1 – 10, where higher scores indicate more fear of the darkside, and was also measured on their willingness to conform to the Empire's demands (Y) on a scale of 1 - 10, where higher scores indicate more willingness to conform. The regression equation and statistics appear in the table below.

Fear of Darkside (X)	Conformity (Y)	Y' = -2.1 + 1.22X
$\overline{X} = 4$	$\overline{Y} = 5$	r = 0.941
SS_X = 50	SS_Y = 84	cov = 6.778

a. Calculate the sum of squares regression and the sum of squares residual.
b. Determine the degrees of freedom regression, degrees of freedom residual, and degrees of freedom total.
c. Calculate the mean square residual and the mean square regression
d. Calculate the F-Ratio.
e. Determine the *p*-value for the F-Ratio.
f. Using α = .05, is the regression equation statistically significant? Why?
g. Calculate the standard error of the estimate from the average residual variance ($MS_{Residual}$).
h. Calculate the standard error of the slope $SE(b_1)$ and standard error of the intercept $SE(b_0)$.
i. Calculate the *t*-Value for the slope (b_1). What is the p-value?
j. Calculate the *t*-Value for the intercept (b_0). What is the p-value?

Chapter 17: Introduction to Analysis of Variance





17.1 Limitations of t-tests

Earlier chapters introduced t-tests that were used to assess the statistical significance of the difference in the means between two levels of an independent variable. The independent-groups t-test was used to compare sample means that differed by levels of a between-subjects independent variable, and the correlated-samples t-test was used to compare sample means that differed by levels of a within-subjects independent variable.

The t-test is used often, because most researchers want to determine which levels of an independent variable are statistically different from each other. However, t-tests have a limit, because you can compare only two levels of an independent variable. The t-test cannot be used to compare three or more levels of an independent variable simultaneously. For example, in a study that examines the ability of *Drug-X* to alleviate anxiety, a researcher may want to examine the effect of taking Drug-X on anxiety compared to taking a placebo and taking nothing. Thus, there are three levels of the independent variable *Drug Condition*: (1) Drug, (2) Placebo, (3) Nothing/Control. How could you analyze these data? To analyze the data, the researcher could conduct several t-tests. With three levels of the independent variable *Drug Condition* there are three pairs of the conditions that can be compared using t-tests:

Pair 1 = Drug vs. Placebo
Pair 2 = Drug vs. Control
Pair 3 = Placebo vs. Control

But there's a problem doing this, because whenever you perform an inferential test on a set of data there is a probability of incorrectly rejecting a true null hypothesis, that is, making a Type I error. The probability of making a Type I error is equal to the chosen or achieved alpha level (α); hence, if you set α = .05, the

probability of making a Type I error is .05. In short, there is always some chance you will be wrong in your statistical decisions (remember, the decisions are based on probabilities).

Importantly, the probability of making a Type I error is additive for all inferential tests performed on a set of data. Specifically, every time you perform a t-test on the same set of data you increase the probability of making a Type I error by the alpha-level for each test. If you set alpha to $\alpha = .05$ for each t-test and you perform one t-test on a set of data, then the probability you will make a Type I error is .05. But if you conduct a second t-test on some part of that same set of data, then the probability you made a Type I error on either test, or both tests, is now .05 + .05 = .10. This is because the probability of making a Type I error was .05 for each test. If you perform a third t-test, then the probability you make at least one Type I error is .15,...,and so on. Thus, for every test you perform on the same data, you increase the chance of making a Type I error by α. This is called **Type I error inflation** and must be avoided.

Because of Type I error inflation it is unwise to conduct multiple t-tests on the same set of data. There is an easy way to ensure you are unlikely to make a Type I error over multiple t-tests, which is accomplished by adjusting the alpha level for all t-tests to be performed. This is the **Bonferroni correction**. In the Bonferroni correction you select an alpha-level that you would like to be at after conducting all necessary t-tests called the **Familywise Type I error rate** (α_{FW}), which should be $\alpha_{FW} = .05$ or less. Then divide the chosen Familywise error rate by the number of tests to be performed. If you plan to conduct three t-tests and you want $\alpha_{FW} = .05$, divide .05 by 3 to get .0167. You use this **adjusted alpha-level** for each t-test, that is, instead of using $\alpha = .05$ for each t-test you use $\alpha = .0167$. After your three t-tests, your overall alpha level will be $\alpha = .05$.

Unfortunately, the Bonferroni correction is very conservative and you may fail to reject a null hypothesis that in other situations you normally would reject. Thus, the Bonferroni is not well-liked. Appendix D discusses the Bonferroni correction as well as several other procedures that are useful for conducting multiple t-tests.

So multiple t-tests are not good a good option when assessing more than two levels of an independent variable...what do you use? We need a test that can simultaneously analyze several levels of an independent variable. Luckily, there is an **Analysis of Variance** (**ANOVA**) procedure that allows one to simultaneously analyze data belonging to two or more levels of an independent variable and also analyze data from several independent variables; thus, this is a very applicable statistical procedure.

17.2 What is ANOVA?

Analysis of variance (ANOVA) does just what its name implies; it *analyzes variance*. ANOVA is preferred over t-tests in situations where you have more than two levels of an independent variable, because you do not need to run several ANOVAs to analyze the data; ANOVA requires just one test. What this means is there is no Type I error inflation when you must compare more than two levels of an independent variable. But, just be aware that you can still produce Type I error inflation if you conducted multiple ANOVAs or other tests on the same set of data.

ANOVA can also be used to analyze data when a research design has several independent variables. Recall that with t-tests you can compare two means drawn from two different samples that are assumed to differ only by the levels of a single independent variable. This is also a limitation of t-tests: The two conditions being compared must differ along one independent variable. With ANOVA you can analyze data from several independent variables simultaneously. For example, in the Drug example in Section 18.1 a researcher could manipulate whether people take Drug-X, Placebo, or Nothing and also manipulate whether people go to a Relaxation Training class or not. If you “cross” the three levels of the *Drug Condition* independent variable with the two levels of the *Therapy* independent variable, you get six combinations, which are represented in the table below:

	Drug Condition

		Drug-X	Placebo	Control
Relaxation Training?	**Yes**	$\overline{X}$	$\overline{X}$	$\overline{X}$
	No	$\overline{X}$	$\overline{X}$	$\overline{X}$

Research designs like these with two or more independent variables are **factorial designs**, and the analyses are known as **factorial analysis of variance**. These will be discussed in Chapter 21. In Chapter 20, the **oneway analysis of variance** is discussed, which analyzes data from conditions that differ by the levels of a single independent variable.

17.3 F-Statistic and Sources of Variance

ANOVA are procedures for analyzing variance, but what 'variance' is being analyzed and where does it come from? I'll use the Drug-X example from above to explain below.

Consider a case in which we manipulate whether different groups of people take Drug-X, a Placebo, or Nothing (Control group). Say there are *n* = 10 people in each group and each person's level of anxiety is measured at the beginning and end of the study and the difference in anxiety between the beginning and end is the dependent variable. Below are hypothetical mean *anxiety difference scores* in each of the three groups (assume that the anxiety difference scores have a range from 0 to 60, with higher values indicating a greater decrease in anxiety):

Drug Condition		
Drug-X	**Placebo**	**Control**
n = 10	n = 10	n = 10
$\overline{X} = 30$	$\overline{X} = 5$	$\overline{X} = 4$

Ignoring the group to which each subject belongs, if the variance in the anxiety scores was calculated across all thirty people you would end up with a measure of **total variation**. That is, all variation in anxiety due to differences among the subjects in the study as well any variation in anxiety across the levels of the independent variable Drug Condition. This total variation includes variation from two sources, that is, between group variance plus the within group variance:

$$\square^2_{\square\square\square\square\square\square} = \square^2_{\square\square\square\square\square\square\square\square} + \square^2_{\square\square\square\square\square\square\square}$$

Between-groups variance is the variation in the dependent variable (anxiety difference scores) between the levels of the independent variable. Thus, between groups variance is seen when there are differences in the means across the levels of the independent variable. Between groups variance is assumed to come from an influence of the independent variable and to sampling error. That is, differences between means in the table above could reflect an **effect** of the independent variable, but could also be due to random sampling **error**. Thus, between group variance is composed of variance due to an effect of the independent variable and to variance due to sampling error:

$$\square^2_{\square\square\square\square\square\square\square\square} = \square^2_{\square\square\square\square\square\square\square} + \square^2_{\square\square\square\square\square\square}$$

Within-groups variance is the variation in the dependent variable (anxiety difference scores) that cannot be attributed to the independent variable. Specifically, it is the variability among scores within each level of the independent variable and summed over all of the levels of the independent variable. Because within-group variance is occurring within levels of the independent variable it does not reflect an influence of the independent variable. Rather, within group variance only reflects random uncontrollable sampling **error** among subjects. Thus, within group variance is composed of only variance due to sampling error:

$$\square^2_{\square\square\square\square\square\square\square} = \square^2_{\square\square\square\square\square\square}$$

Remember, between-group variance reflects an influence of the independent variable on the dependent variable and sampling error, whereas within-group variance reflects only sampling error. If one could show that variation between groups is greater than the random variation within groups (sampling error), it could reflect a statistically significant influence of the independent variable on the dependent variable. How do you determine whether the variability due to the effect of the independent variable is greater than the random sampling error? Divide the between group variance by the within group variance, which forms a variance ratio called the **F-Ratio**:

$$F = \frac{\sigma^2_{Between}}{\sigma^2_{Within}}$$

Between group variance reflects both the effect of the independent variable and random error; whereas within group variance reflects only sampling error; thus:

$$F = \frac{(\sigma^2_{Effect} + \sigma^2_{Error})}{\sigma^2_{Error}}$$

By dividing a value that represents the combination of effect and error by error alone, the F-Ratio reflects the variability due to the independent variable. Importantly, the expected value of the F-Ratio given the null hypothesis is true and there is no effect of the independent variable is equal to 1 [i.e., $E(F \mid H_0 = True) = 1$]. Hence, if there is no influence of the independent variable on the dependent variable, the F-Ratio should be a ratio of sampling error over sampling error; and dividing a value by itself yields a value of 1. Thus, the expected F-Value under the null is one and F-Ratio can never be equal to zero, or negative. In practice, because we are basing our analyses on sample data we need to estimate the between group variance and the within groups variance. These estimated variances are called **mean squares** (**MS**) and are obtained by dividing a sum of squares by a degree of freedom value:

$$MS = \frac{SS}{df}$$

Because we have to estimate two sources of variance (between groups and within groups), we calculate a **mean square within groups** (**MS_W**) and a **mean square between groups** (**MS_B**):

$$MS = \frac{SS_B}{df_B} \quad \text{and} \quad MS = \frac{SS_W}{df_W}$$

The mean square within groups and mean square between groups form the F-Ratio:

$$F = \frac{MS_B}{MS_W}$$

Thus, instead of relying on population parameters for the F-Ratio, we estimate the parameters from sample data, which we will go over in detail in the following chapters.

17.4 Hypotheses with ANOVA

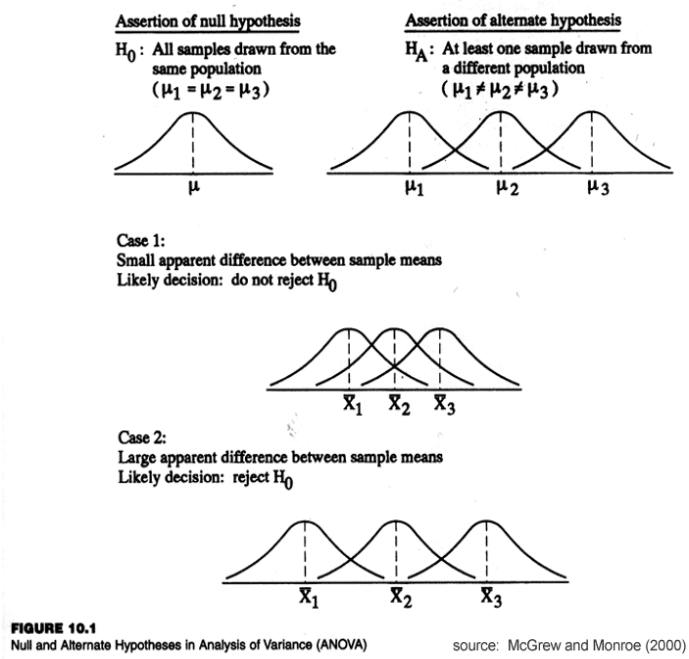


FIGURE 10.1
Null and Alternate Hypotheses in Analysis of Variance (ANOVA) source: McGrew and Monroe (2000)

Hypotheses with ANOVA can get cumbersome to write especially with more than one independent variable. This is why most researchers set all of the means equal to each other under the null hypothesis and set all of the means as unequal to each other under the alternative hypothesis. Using the example from above the hypotheses are:

$$H_0: \mu_{Drug\text{-}X} = \mu_{Placebo} = \mu_{Control}$$
$$H_1: \mu_{Drug\text{-}X} \neq \mu_{Placebo} \neq \mu_{Control}$$

The null hypothesis states that all three means are predicted to equal one another (no difference between means); and the alternate hypothesis states that there will be some difference between the means, that is, some influence of the independent variable. Note that the alternative hypothesis is not stating which pairs of means are predicted to be different; the difference could be between the Drug-X and Placebo groups, between the Drug-X and Control groups, between the Placebo and Control groups, or between all groups.

17.5 Experimental Designs in ANOVA

What types of research designs can be analyzed with an analysis of variance? There are ANOVA procedures for analyzing between-subjects experimental designs, ANOVA procedures for analyzing within-subjects experimental designs, and procedures analyzing crossed or mixed-model designs. These latter research designs include at least one variable manipulated between-subjects and at least one variable manipulated within-subjects. Such designs are mixed in the sense that that include a mixture of within-subject and between-subject independent variables.

17.6 Subscripts

I have resisted using subscripts and superscripts in the statistical formulas to this point, because they generally make things more confusing than they need to be, but for ANOVA they are necessary. Before

moving onto actual ANOVA procedures I want to address a little bit of the nomenclature you will encounter in Chapters 19 and 20. The formula below is used for obtaining something called the **within-group sums of squares**, which is a measure of the total within group variation in a set of data:

$$SS_{WG} = \sum_{k=1}^{K} \sum_{i=1}^{n} \left(X_{ik} - \bar{X}_k\right)^2$$

Yes, there two summation symbols and no, this is not a mistake. This formula is really not all that different from the sums of squares formula used throughout the book. The superscript 'K' refers to the levels of an independent variable that are being compared. In the example above there are three levels of the independent variable *Drug condition*. The subscript 'k' refers to the k^{th} level of that independent variable. The superscript 'n' refers to the number of subjects included in each level of the independent variable. In the example above, there were n = 10 subjects in each level. The subscript 'i' refers to the i^{th} person in that level of the independent variable. For example, if k = 1 and i = 1, as above, this refers to the first subject in the first level of the independent variable. If k = 2 and i = 3, this refers to the 3^{rd} subject in the 2^{nd} level of the independent variable. And so on.

Consider first the portion of the formula in the parentheses. The X_{ik} refers to the dependent variable score from individual 'i' from level 'k', that is, the score of some individual who belongs to some level of the independent variable. The $\bar{X}_k$ value refers to the mean of the level from which that individual (X_{ik}) came. For example, if you have Drug, Placebo, and Control groups as the three levels of an independent variable and there are n = 10 subjects in each group: X_{ik} may refer to the score of the 4^{th} subject from the Placebo group, and $\bar{X}_k$ would refer to the mean of the Placebo group. Or, X_{ik} could be the 9^{th} subject in the Control group, and $\bar{X}_k$ would refer be the mean of the Control group. Hence, the part in the parentheses is simply telling you to calculate the difference between an individual's dependent variable score and the mean of his/her group.

The summation symbol with the superscript n tells you to do this for each individual in a particular level of the independent variable. That is, calculate the difference between each individual's score and the mean of that group, square those differences, and sum the squared differences. The summation with the superscript K tells you to do this for each level of the independent variable. That is, after you find the sum of squares for one level, calculate the sum of squares for each other level. Thus, all that this equation above is telling you to do is find the deviation between an individual's score and the mean of that individual's group, square that difference, sum the squared differences, and do this for each subject in each group. The subscripts and nomenclature look crazy at first, but the best to do is to identity what 'i', 'n', 'k', and any other nomenclature stand for. Once you have that figured out, start within the parentheses and work your way out from there.

CH 17 Homework Questions

1. What does the alternate hypothesis ask for a one-way between-subjects analysis of variance?

2. What is the difference between between-group variability and within-group variability?

3. What does between-group variability reflect? Where does it come from?

4. What does within-group variability reflect? Where does it come from?

5. When will the *F* ratio approach 1.00? When will the *F* ratio be greater than 1.00?

6. What is the difference between the sums of squares total, the sums of squares between-groups, and the sums of squares within-groups? How are they related?

7. Consider the scores in an experiment involving four levels of an independent variable. Without computing anything, what value must the sums of squares within-groups equal? Why?

A	B	C	D
10	1 2	9	4
10	1 2	9	4
10	1 2	9	4
10	1 2	9	4
10	1 2	9	4

8. State the critical value of *F*-Value for a oneway ANOVA under each of the following conditions. Note the n values are the number of subjects in each group (they are not n_T):

a. k = 3, n = 7, α = .01
b. k = 5, n =10, α = .01
c. k = 4, n = 5, α = .05
d. k = 6, n = 4, α = .05
e. k = 3, n = 10, α = .01
f. k = 2, n = 12, α = .05
g. k = 5, n = 15, α = .01
h. k = 4, n = 5, α = .05

9. Determine the p-values for a oneway ANOVA under each of the following conditions. Note the n values are the number of subjects in each group (they are not n_T):

a. k = 2, n = 10, *F* = 2.50
b. k = 3, n = 10, *F* = 2.50
c. k = 2, n = 8, *F* = 5.20
d. k = 4, n = 8, *F* = 5.00
e. k = 3, n = 6, *F* = 4.20
f. k = 3, n = 12, *F* = 4.20

10. State the null and alternate hypotheses for a oneway ANOVA with 5 levels of the independent variable:

11. Insert the missing entries in the summary table for a one-way analysis of variance having four levels of the independent variable.

Source	SS	df	MS	F
Between		3	18	3.5
Within				
Total		23		

12. Insert the missing entries in the summary table for a one-way analysis of variance having three levels of the independent variable and *n* = 20.

Source	SS	df	MS	F
Between			25	
Within	140			
Total	190	99		

13. Insert the missing entries in the summary table for a one-way analysis of variance having three levels of the independent variable and $n = 15$.

Source	SS	df	MS	F
Between		2		10
Within			4	
Total	128			

14. Insert the missing entries in the summary table for a one-way analysis of variance having seven levels of the independent variable and $n = 32$.

Source	SS	df	MS	F
Between		6	5125	
Within	2500	25		
Total				

15. Assume you have an independent variable, A, with three levels (A_1, A_2, A_3). If you set $\alpha = .05$, what is the Familywise Type I error rate after running all possible t-tests?

16. Assume you have an independent variable, A, with four levels (A_1, A_2, A_3, A_4). If you set $\alpha = .01$, what is the Familywise Type I error rate after running all possible t-tests?

17. Assume you have an independent variable, A, with four levels (A_1, A_2, A_3, A_4). If you set $\alpha = .03$, what is the Familywise Type I error rate after running all possible t-tests?

18. Assume you have an independent variable, A, with four levels (A_1, A_2, A_3, A_4). If you want your Familywise Type I error rate to be .05 after running all possible t-tests, what must the alpha level be set to for each test?

19. Assume you have an independent variable, A, with three levels (A_1, A_2, A_3). If you want your Familywise Type I error rate to be .02 after running all possible t-tests, what must the alpha level be set to for each test?

Chapter 18: Oneway Between Subjects ANOVA

18.1 What is The 'Oneway' Design?

The simplest research design that can utilize analysis of variance includes one independent variable with at least two levels. In such designs, it does not matter whether you use a t-test or ANOVA and, indeed, if you performed an ANOVA and then performed a t-test on the same data, you would find that $F = t^2$.

ANOVA conducted on data from a design with one independent variable is known a **oneway ANOVA**, where *oneway* refers to the one independent variable. This chapter examines how to properly conduct oneway between-subjects ANOVA, that is, analyze the influence of a single independent variable that was manipulated between-subjects. Oneway ANOVA is used to assess the relationship between variables when:

1. The dependent variables is *quantitative*
2. The independent variable is *qualitative* in nature
3. The independent variable has two or more levels
4. The independent variable is manipulated between-subjects

For example, say an educational psychologist is interested in the influence of testing location on test performance. Specifically, does the location where you study for a test have an influence on test scores? The researcher samples $n = 12$ students and has each student study a list of 20 words. One week later each student's memory for the words is tested by having each student recall as many studied words as possible. The manipulation is the testing location.

One group of $n = 3$ students is tested in the same room where the words were studied (*Same* condition); a second group of $n = 3$ students is tested in a different room, but with the same characteristics as the original room (*Different* condition); a third group of $n = 3$ students is tested in a basement (*Basement* condition); and a fourth group of $n = 3$ students is tested in the basement, but is presented with a picture of the room where the words were studied (*Picture* condition). Below, the number of words correctly recalled by each student is presented for each group:

Condition							
Same		**Different**		**Basement**		**Picture**	
Subject	**X_S**	**Subject**	**X_D**	**Subject**	**X_B**	**Subject**	**X_P**
A	8	D	5	G	3	J	7
B	10	E	7	H	5	K	8
C	6	F	6	I	4	L	6
$\sum X_S = 24$		$\sum X_D = 18$		$\sum X_B = 12$		$\sum X_P = 21$	
$M_S = 8$		$M_D = 6$		$M_B = 4$		$M_P = 7$	

Because there are more than two levels of the independent variable, these data will be analyzed using a between-subjects oneway ANOVA. The prediction is that the location in which the recall test was taken will have some effect on test performance. Thus, the null and alternate hypotheses are below and the alpha level will be set to $\alpha = .05$:

$$H_0: \mu_S = \mu_D = \mu_B = \mu_P \qquad H_1: \mu_S \neq \mu_D \neq \mu_B \neq \mu_P$$

18.2 Numeric Example

Before beginning, it is good to construct an **ANOVA summary table** (see example at right), in which you list each degrees of freedom (df), sum of squares (SS), mean square (MS), and the F-Ratio, as each value is calculated. Remember from Chapter 17 there are three sources of variance: between-group variance, within-group variance, and total variance. Hence, we need to calculate three sums of squares and three degrees of freedom values, and then find two *mean squares* and the F-Ratio.

Variance Source	SS	df	MS	F
Between	??	??	??	??
Within	??	??	??	
Total	??	??		

Between-subjects degrees of freedom (**df_B**) are equal to k – 1, where k is the number of levels of the independent variable. In this case there are four levels of the independent variable (Same, Different, Basement, and Picture); thus, k = 4, and $df_B = 4 - 1 = 3$. **Within subjects degrees of freedom** (**df_W**) are equal to n_T – k, where n_T is the total number of subjects. In this case there are $n_T = 12$ subjects, so $df_W = 12 - 4 = 8$. Another way to obtain degrees of freedom within subjects is multiplying the degrees of freedom per level of the independent variable (n – 1) by the number of levels of the independent variable (k), such that $df_W = k(n - 1)$. In this example, there are n = 3 subjects in each level of the independent variable; thus, $df_W = 4(3 - 1) = 8$. (This works only when there are an equal number of subjects in each level of the independent variable.) **Total degrees of freedom** (**df_T**) are equal to n_T – 1. In this example, with twelve subjects, $df_T = 12 - 1 = 11$. Notice that $df_T = df_B + df_W$ (11 = 3 + 8).

Next, calculate the sum of squares for each of the three sources of variance starting with sum of squares of total variance; the **sum of squares total** (**SS_T**). The sum of squares total is the total variation over all scores in the data and is found by comparing each individual value to the **grand mean** of the data. The grand mean ($\overline{\overline{X}}$ or G) is the mean of all the scores:

$$\overline{\overline{X}} = G = \frac{\sum X}{n} = \frac{75}{12} = 6.25$$

The formula for the **sum of squares total** is:

$$SS_T = \sum_{k=1}^{K} \sum_{i=1}^{n} (X_{ik} - G)^2$$

To find sum of squares total: subtract the grand mean from each score (X_{ik}). Next, square each difference, and then sum the squared differences. This is demonstrated in the table below. As you can see, calculating total sum of squares is no different than calculating a sum of squares for a sample.

Subject	X_{ik}	$(X_{ik} - G)$	$(X_{ik} - G)^2$
A	8	1.750	3.062
B	10	3.750	11.062
C	6	-0.250	0.062
D	5	-1.250	1.562
E	7	0.750	0.562
F	6	-0.250	0.062
G	3	-3.250	10.562
H	5	-1.250	1.562
I	4	-2.250	5.062
J	7	0.750	0.562
K	8	1.750	3.062
L	6	-0.250	0.062
			$SS_T = 40.244$

Next, calculate **sum of squares within groups** (**SS_W**). The formula for within-group sum of squares is:

$$SS_W = \sum_{k=1}^{K} \sum_{i=1}^{n} (X_{ik} - \bar{X}_k)^2$$

This look similar to the formula for SS_T, except the grand mean has been replaced by the mean of each group. When calculating the sum of squares within groups you are actually finding the sum of squares for each level of the independent variable and adding those sums of squares across all levels of the independent variable. Specifically, first you calculate the difference between an individual's score (X_{ik}) and the mean of the level of the independent variable that individual belongs. You do this for each individual, being careful to subtract the correct mean from each individual. Next, square these differences and add them to get SS_W. This is demonstrated in the table below, where the values for each individual are included as a reference:

Subject	**X_{ik}**	$\bar{X}_k$	$(X_{ik} - \bar{X}_k)$	$(X_{ik} - \bar{X}_k)^2$
A	8	8	8 - 8 = 0	0
B	10	8	10 - 8 = 2	4
C	6	8	6 - 8 = -2	4
D	5	6	5 - 6 = -1	1
E	7	6	7 - 6 = 1	1
F	6	6	6 - 6 = 0	0
G	3	4	3 - 4 = -1	1
H	5	4	5 - 4 = 1	1
I	4	4	4 - 4 = 0	0
J	7	7	7 - 7 = 0	0
K	8	7	8 - 7 = 1	1
L	6	7	6 - 7 = -1	1
				$SS_W = 14$

The sum of squares within groups represents the summed variability within each level of the independent variable and summed across all levels. Specifically, it is the sum of squares for each level of the independent variable, added together.

Finally, calculate the sum of **squares between groups** (SS_B):

$$SS_B = \sum_{k=1}^{K} n_k (\bar{X}_k - G)^2$$

This formula is a bit different from the previous two, but is just another sum of squares. The part of the formula in parentheses tells you to subtract the grand mean from the mean associated with a level of the independent variable. You next square this mean difference and then multiply that squared mean difference by the sample size for that level of the independent variable (n_k). Once this is done for each group, add the products together to get SS_B. Remember, n_k is equal to 3 for each level of the independent variable in this example. This is demonstrated in the table below:

Group	$\bar{X}_k$	$(\bar{X}_k - \bar{\bar{X}})$	$(\bar{X}_k - \bar{\bar{X}})^2$	$n_k(\bar{X}_k - \bar{\bar{X}})^2$
Same	8	1.750	3.062	9.186
Different	6	-0.250	0.062	0.186
Basement	4	-2.250	5.0620	15.186
Picture	7	0.750	0.562	1.686
				$SS_B = 26.244$

This value, (SS_B = 26.244) is a measure of the total variability between the means of the levels of the independent variable. We now have SS_T, SS_W, and SS_B. Note that $SS_T = SS_W + SS_B$ (40.244 = 14 + 26.244). In practice, you could actually calculate two of the three sums of squares values and use those two to obtain the third value.

Next, calculate the **mean squares** (**MS**) that will then be used to calculate the F-Ratio; specifically, a **mean-square between groups** (**MS_B**) and a **mean-square within groups** (**MS_W**). Calculating each mean square simply involves dividing a sum of squares by associated degrees of freedom. Remember, the mean square is the average variability for a source of variance. Mean-square between groups is:

$$MS_B = \frac{SS_B}{df_B} = \frac{26.244}{3} = 8.748$$

Mean square within groups is:

$$MS_W = \frac{SS_W}{df_W} = \frac{14}{8} = 1.75$$

The F-Ratio is calculated by dividing the mean square between groups (effect) by the mean square within groups (error):

$$F = \frac{MS_B}{MS_W} = \frac{8.748}{1.75} = 5$$

The table to the right is the complete ANOVA summary table is listed with each of the values we just calculated. The value of F = 5 is the obtained F-Ratio, which is used to determine whether the analysis of variance was statistically significant.

Variance Source	SS	df	MS	F
Between	26.244	3	8.748	5
Within	14	8	1.75	
Total	40.244	11		

To assess the significance of the F-Ratio you look-up a *p*-value in Table 3 in Appendix A. Table 3 (there are six of them) lists the probabilities associated with areas under the F-Distribution. Each of the six tables 3 are associated with a different number of degrees of freedom between groups (df_B), ranging from $df_B = 1$ through $df_B = 10$ (see upper left corner of each table). Each table is structured the same, with F-Ratio values in the leftmost column (F) and degrees of freedom error/within subjects (df_W) values in each column heading. In the example there are $df_B = 3$, so we'll use Table 3-C, a portion which is reproduced, below.

	Degrees of Freedom Error (df_W)																										
F	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**14**	**16**	**18**	**20**	**22**	**24**	**26**	**28**	**30**	**40**	**45**	**50**	**60**	**80**	**100**	**250**	**1000**
4.50	0.1872	0.1242	0.0901	0.0695	0.0558	0.0464	0.0395	0.0343	0.0303	0.0246	0.0207	0.0178	0.0156	0.0139	0.0125	0.0114	0.0104	0.0096	0.0089	0.0071	0.0066	0.0062	0.0056	0.0049	0.0045	0.0041	0.0038
4.60	0.1837	0.1210	0.0873	0.0668	0.0535	0.0442	0.0375	0.0325	0.0286	0.0232	0.0194	0.0167	0.0146	0.0130	0.0117	0.0106	0.0097	0.0089	0.0083	0.0066	0.0061	0.0057	0.0052	0.0045	0.0041	0.0037	0.0034
4.70	0.1803	0.1180	0.0845	0.0642	0.0512	0.0421	0.0356	0.0307	0.0269	0.0217	0.0181	0.0156	0.0136	0.0121	0.0108	0.0098	0.0089	0.0082	0.0076	0.0060	0.0056	0.0052	0.0047	0.0041	0.0037	0.0034	0.0031
4.80	0.1770	0.1152	0.0819	0.0620	0.0491	0.0402	0.0338	0.0290	0.0253	0.0203	0.0168	0.0144	0.0125	0.0111	0.0099	0.0089	0.0081	0.0074	0.0068	0.0054	0.0050	0.0046	0.0042	0.0036	0.0033	0.0030	0.0028
4.90	0.1738	0.1125	0.0794	0.0598	0.0471	0.0384	0.0321	0.0274	0.0238	0.0189	0.0156	0.0133	0.0115	0.0101	0.0090	0.0081	0.0073	0.0066	0.0061	0.0048	0.0045	0.0041	0.0037	0.0032	0.0029	0.0026	0.0025
5.00	0.1710	0.1096	0.0770	0.0577	0.0452	0.0367	0.0306	0.0260	0.0224	0.0176	0.0144	0.0122	0.0105	0.0092	0.0081	0.0073	0.0066	0.0060	0.0055	0.0043	0.0040	0.0037	0.0033	0.0029	0.0026	0.0023	0.0022
5.10	0.1683	0.1070	0.0747	0.0557	0.0434	0.0351	0.0291	0.0247	0.0212	0.0165	0.0134	0.0113	0.0097	0.0085	0.0075	0.0067	0.0060	0.0055	0.0050	0.0039	0.0036	0.0033	0.0030	0.0026	0.0023	0.0021	0.0019
5.20	0.1655	0.1045	0.0726	0.0537	0.0417	0.0335	0.0277	0.0234	0.0201	0.0155	0.0125	0.0105	0.0090	0.0079	0.0069	0.0062	0.0056	0.0051	0.0046	0.0036	0.0033	0.0030	0.0027	0.0023	0.0021	0.0019	0.0017
5.30	0.1628	0.1021	0.0705	0.0519	0.0401	0.0320	0.0264	0.0222	0.0190	0.0146	0.0117	0.0098	0.0084	0.0073	0.0064	0.0057	0.0052	0.0047	0.0043	0.0033	0.0030	0.0027	0.0025	0.0021	0.0019	0.0017	0.0015
5.40	0.1602	0.0998	0.0685	0.0502	0.0385	0.0305	0.0251	0.0211	0.0180	0.0137	0.0109	0.0092	0.0078	0.0068	0.0059	0.0053	0.0048	0.0043	0.0040	0.0030	0.0028	0.0025	0.0023	0.0020	0.0017	0.0015	0.0014
5.50	0.1577	0.0976	0.0666	0.0485	0.0371	0.0291	0.0240	0.0201	0.0171	0.0130	0.0103	0.0086	0.0073	0.0063	0.0055	0.0049	0.0044	0.0040	0.0037	0.0028	0.0025	0.0023	0.0021	0.0018	0.0016	0.0014	0.0013

To locate the p-value associated with the obtained F-Ratio (5.00), locate the $df_W = 8$ value in the column headings (highlighted in yellow, above) and scroll down that column until you come to the row associated with F = 5.00 (highlighted in yellow, above). The p-value associated with this F-Ratio is .0306, which is less than the selected alpha level of .05. This means that the outcome of the ANOVA is statistically significant. In this case, we reject the null hypothesis and accept the alternate hypothesis. This significant difference in an oneway ANOVA is known as a **main effect**, which means there is a significant difference among the means along independent variable.

18.3 Post Hoc Testing

A statistically significant ANOVA tells you the independent variable had some influence on the dependent variable, and that at least two of the means were significantly different from each other; however,. the ANOVA does not tell which means are significantly different. To determine which means are significantly different from each other we need to run a **post-hoc test**. Post-hoc tests are used to determine which means are significantly different from each other. It is important to remember that post-hoc testing should be conducted only if an ANOVA was significant; if the ANOVA is not statistically significant you should never do post-hoc testing, because any results from those post-hoc tests would be meaningless.

A common post-hoc test is **Tukey's Honestly Significant Difference** (**HSD**). The value calculated in the Tukey's test is a *common difference* (**CD**) that must be found between means to conclude the difference between means is statistically significant. That is, we'll calculate the Tukey's HSD value and then compare each pair of means to find each mean difference. If the difference between any two means exceeds the Tukey's HSD value, then that difference is considered to be statistically significant. The formula for Tukey's HSD is:

$$HSD = q\sqrt{\frac{MS_W}{n_k}}$$

In the formula, MS_W is the *mean-square within groups* calculated in the ANOVA above ($MS_W = 1.75$), and n_k is the number of subjects per level of the independent variable; that is, the number of subjects in each level of the independent variable ($n_k = 3$). Note, n_k must be equal across the levels of the independent variable to use this version of the Tukey's test.

The q is the **Studentized range statistic** and is obtained from Table 4 in Appendix A. To find the value for q, in the column headings, look up the number of levels of the independent variable that were being compared in the ANOVA, which in the example was k = 4. Next, move down that column until you come to the rows associated with df_W from the ANOVA, which was $df_W = 8$ in the example above. You will see two values, 4.53 and 6.19. Use the upper value (q = 4.53) if α = .05 in the ANOVA, and use the lower value (q = 6.20) if α = .01 in the ANOVA. Because α =.05 in the ANOVA above, we'll use the q = 4.53.
Tukey's HSD for this example is:

$$HSD = 4.53\sqrt{\frac{1.75}{3}} = 3.461$$

This value (3.461) is the minimum difference needed for a difference between means to be considered significantly significant. Looking at that the means for the four groups in the example we have:

Same condition:	$M_S = 8$	Different condition:	$M_D = 6$
Basement condition:	$M_B = 4$	Picture condition:	$M_P = 7$

With HSD = 3.461, the only two means that are statistically significant from each other is the difference between the Same condition and Basement condition, which has a mean difference of 4. Note that when comparing means, you are just looking for the absolute difference between the means.

Unfortunately, this version of the Tukey's HSD test cannot be used when you have unequal sample sizes across the levels of the independent variable. If you have unequal sample sizes across the level of the independent variable, you can use **Fisher's Least Significant Difference** (**LSD**) test as a post-hoc test following a statistically significant ANOVA test.

The logic of Fisher's LSD is the same as Tukey's HSD. First calculate an LSD value that represents the minimum differences that is needed between a pair of means for those means to be considered statistically significant. The catch is that with the Fisher's LSD test, you have to calculate a new LSD value for each pair of means, due to unequal sample sizes across levels of the independent variable. But, if you have equal sample sizes across the levels of your independent variable, you only need to calculate one LSD value. Hence, in our example in this chapter, we need to calculate only one Fisher's LSD value. Fisher's LSD formula is:

$$LSD = t_{\alpha}\sqrt{MS_W(\frac{1}{n_1})(\frac{1}{n_2})}$$

The MS_W is the mean square within subjects from the ANOVA. The n_1 and n_2 values are the sample sizes of the two groups that are being compared in the Fisher's LSD test. In the present example, because all of the sample sizes are equal to $n_k = 3$, $n_1 = n_2 = 3$, so we only need to calculate on LSD value. But, if we had two samples with unequal sizes, then n1 ≠ n2 and we might have to calculate several Fisher's LSD values. The t_{α} value is the critical t-Value that would be used if comparing these two groups with an independent groups t-test. This t-Value is obtained by looking up a critical t-Value for a non-directional hypothesis in Table 2 in Appendix A or in the statistics packet, using the alpha level from the ANOVA, and the degrees of freedom for comparing these two levels of the independent variable ($df = n_1 + n_2 – 2$). In this example $\alpha = .05$ from the ANOVA above and $df = 3 + 3 – 2 = 4$. Thus, $t_{\alpha} = 2.78$. Using this information, the value of Fisher's LSD is:

$$LSD = 2.78\sqrt{1.75(\frac{1}{3})(\frac{1}{3})} = 3.002$$

Use this Fisher's LSD value to determine whether the means from the example above are significantly different from each other. Again, you would normally calculate a different Fisher's LSD any time you are comparing two levels of the independent variable that have different sample sizes, but because we have equal sample sizes we only need to calculate the one LSD value. Also, note that Fisher's LSD = 2.998 is smaller than the Tukey's HSD = 3.461 that we calculated above. This will almost always be the case, which makes Fisher's LSD a more powerful post-hoc test than Tukey's HSD. In the present example, the mean difference between the Same group (*M* = 8) and the Basement group (*M* = 4) is still statistically significant, because the mean difference (4) is greater than the LSD value. But now, the mean difference between the Picture (*M* = 7) group and Basement group (*M* = 4) is nearly significant. Hence, you are more likely to say two groups are significantly different with Fisher's test.

18.4 Effect Size and Explained Variance

Like other inferential statistics you can also calculate the effect size (η^2) of the significant main effect; that is, the proportion of the total variance that is explained in the relationship between the independent variable and dependent variable. In an analysis of variance the total variance is equal to SS_T. The relationship between the independent variable and dependent variable. The effect of the independent variable on the dependent variable is the variation between groups (SS_B). If you divide the between-groups variation (SS_B) by total variation (SS_T) you are left with a value that is equal to the proportion of variance explained in the relationship between the independent variable and dependent variable. Thus, from the example above, effect size is:

$$\eta^2 = \frac{SS_B}{SS_T} = \frac{26.244}{40.244} = 0.652$$

About 65% of the variance in the dependent variable is attributable to the independent variable. Another popular measure of effect size is **Cohen's *f***, which is similar to Cohen's d from earlier chapters. Cohen's *f* can be conceptualized as the standard deviation of the 'effect' over the standard deviation of the 'error':

$$f = \frac{\sigma_{Effect}}{\sigma_{Error}}$$

However, it is often not clear how the standard deviation of the 'effect' and the standard deviation of the 'error' can be calculated. Thus, most researchers opt for calculating Cohen's f from the eta-squared value:

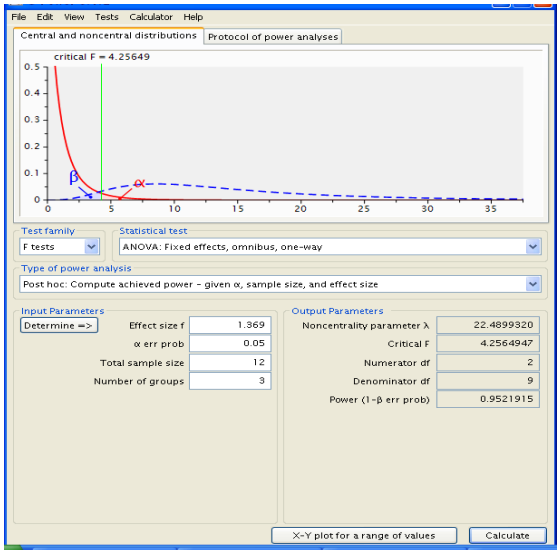
$$f = \sqrt{\frac{\eta^2}{1-\eta^2}} = \sqrt{\frac{0.652}{1-0.652}} = 1.369$$

Cohen also provides useful labels to describe how 'big' or how 'strong' of a relationship the effect size indicates. The table below reports the minimum Cohen's d and eta-squared values that correspond to 'weak', 'moderate', and 'strong' effect sizes (also called 'small', 'medium', and 'large' effect sizes):

Effect Size	**Cohen's *f***	**η^2**
"Small" ("Weak")	.10	.01
"Medium" ("Moderate")	.25	.06
"Large" ("Strong")	.50	.14

18.5 Statistical Power for Oneway ANOVA

Recall, from earlier chapters, statistical power (1 - β) is the probability of correctly rejecting a false null hypothesis. You want this probability to be high, so that a statistically significant result most likely reflects a true significant difference; or to show that a non-significant difference is most likely not significant. Thus, you generally want the statistical power of an inferential test to be .80 or greater. Using the ANOVA example from earlier sections, we can determine our achieved power in the analysis of variance for correctly rejecting a false null hypothesis. Recall, that the total number of subjects in the data set was n = 12, the alpha level was α = .05, and the effect size was f = 1.369.



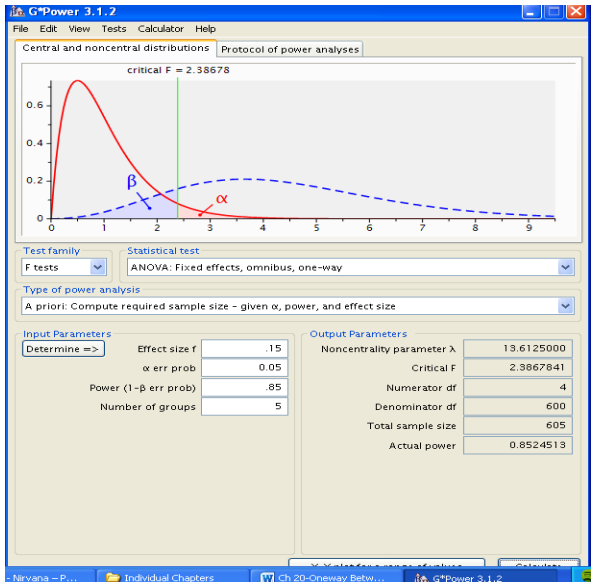
To perform a post hoc power analysis, open G*Power 3 and, under "Test family," select "F tests." Under "Statistical test" choose "ANOVA: Fixed effects, omnibus, one-way", and under "Type of Power Analysis" choose "Post hoc: Compute achieved power." Next, you'll need to enter the α-level, total sample size, effect size and the number of groups from the t-test above. The alpha level was α = .05, *n* = 12, *f* = 1.369, and number of groups was three (k = 3).

After clicking calculate, G*Power tells you that the statistical power is .9521915, which is high and is good. Thus, there is a high probability that we are correctly rejecting the null hypothesis if it is indeed false.

We can also use G*Power in a priori power analyses for a oneway ANOVA to determine the number of subjects needed to end up with a desired level of power. For example, assume I am planning to collect data from a sample of subjects in a between subjects design that has five levels of the independent variable (k

= 5). I know from prior studies that the effect size for such a manipulation is generally “small” and near *f* = .15. For my inferential test I am planning to use α = .05 and I want Power = .85.

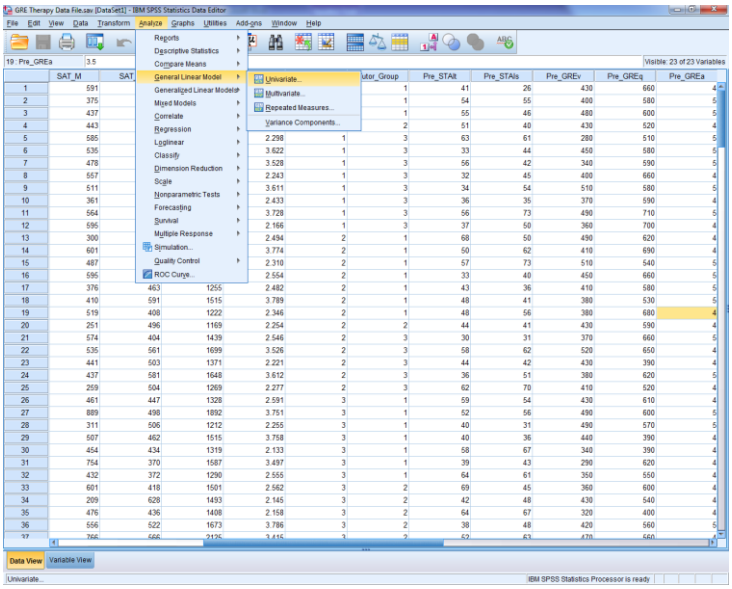
To determine the appropriate sample size, open G*Power 3 and under “Test family” be sure “F tests” is chosen, under “Statistical test” choose “ANOVA: Fixed effects, omnibus, one-way”; and under “Type of Power Analysis” choose “A priori: compute required sample size.” Next, you'll need to enter the α-level, expected effect size, desired power, and number of groups. After clicking Calculate you are provided with the total sample size needed to achieve this amount of Power, which is *n* = 605.



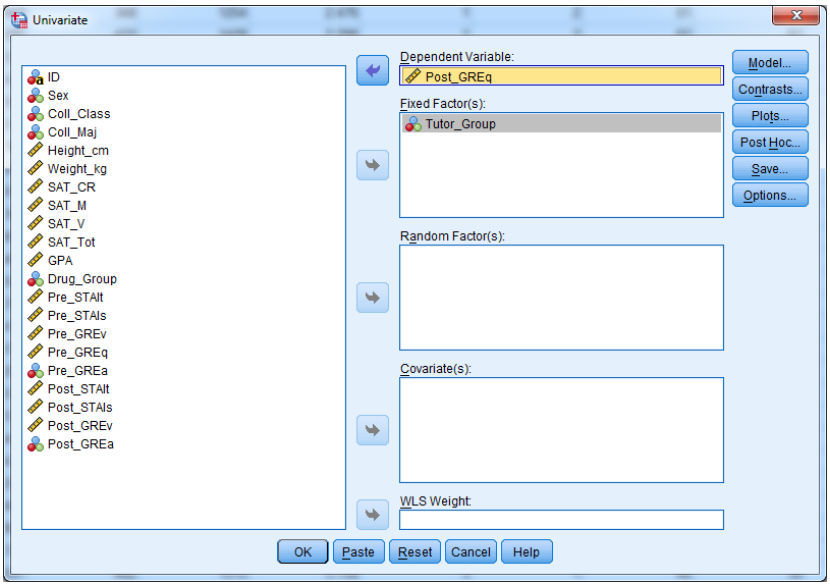
18.6 Oneway ANOVA in SPSS

The following uses the GRE Therapy Data file. Recall, this data file is based on a hypothetical study examining the influences of a study-aid drug and types of tutoring on performance on the Graduate Record Examinations (GREs). The data file includes a number of dependent variables and as well as the two independent variables (Drug_Group and Tutor_Group). In particular, GRE scores were obtained *prior* to introducing the independent variables (Pre_GREv, Pre_GREq, Pre_GREa) and *after* introducing the independent variables (Post_GREv, Post_GREq, Post_GREa).

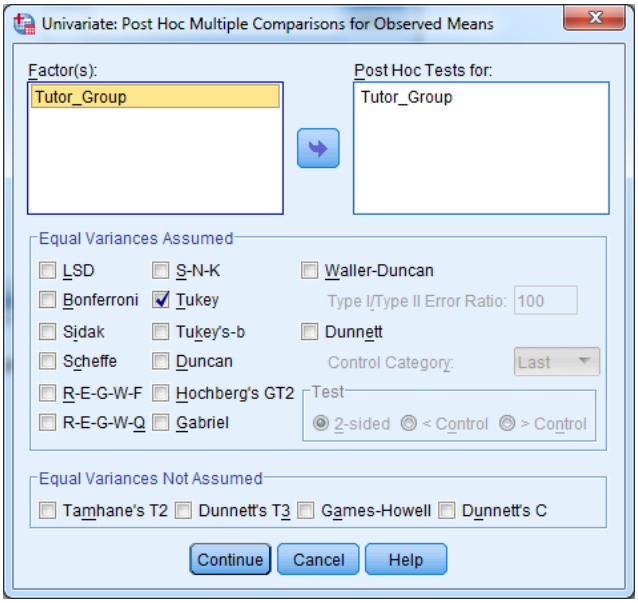
Let’s say we want to know whether the post-test (after the independent variables were introduced) GRE Quantitative Scores (Post_GREq) differ across the levels of the independent variable Tutor_Group. In this data set, there are three levels of the variable Tutor Group: No Tutoring, Group Tutoring, and Individual Tutoring. To request SPSS perform an ANOVA, from the Analyze menu, select General Linear Model, and the Univariate:



In the new window that opens, we need to declare the dependent variable (Post_GREq), which goes in the area labeled Dependent Variable, and also declare the independent variable (Tutor_Group), which goes in the area labeled Fixed Factor(s). To do this, move the variables from the area on the left to the appropriate areas on the right:



Before running the ANOVA, we need to request post-hoc tests and descriptive statistics. First, click on the 'Options…' button and make sure Descriptives is checked. Next, click the 'Post Hoc…' button. Highlight the Tutor_Group actor on the left and click the arrow button; this tells SPSS you want to compare the levels of that independent variable. There are many pot-hoc tests to choose from, but in this example, be sure Tukey is checked off, and then click OK.



In the main window, click 'OK' and the ANOVA will be conducted. You should receive output resembling that on the next page. Comments on the meaning of each table appear on the following page.

Univariate Analysis of Variance

Between-Subjects Factors

		Value Label	N
Tutor_Group	1	Control Group (no tutoring)	80
	2	Group Tutoring	80
	3	Individual Tutoring	80

Descriptive Statistics

Dependent Variable: Post_GREq

Tutor_Group	Mean	Std. Deviation	N
Control Group (no tutoring)	574.25	76.502	80
Group Tutoring	589.00	74.436	80
Individual Tutoring	610.13	84.606	80
Total	591.13	79.685	240

Tests of Between-Subjects Effects

Dependent Variable: Post_GREq

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	52022.500[a]	2	26011.250	4.206	.016
Intercept	83862903.750	1	83862903.750	13561.589	.000
Tutor_Group	52022.500	2	26011.250	4.206	.016
Error	1465573.750	237	6183.855		
Total	85380500.000	240			
Corrected Total	1517596.250	239			

a. R Squared = .034 (Adjusted R Squared = .026)

Post Hoc Tests

Tutor_Group

Multiple Comparisons

Dependent Variable: Post_GREq
Tukey HSD

(I) Tutor_Group	(J) Tutor_Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Control Group (no tutoring)	Group Tutoring	-14.75	12.434	.463	-44.08	14.58
	Individual Tutoring	-35.88*	12.434	.012	-65.20	-6.55
Group Tutoring	Control Group (no tutoring)	14.75	12.434	.463	-14.58	44.08
	Individual Tutoring	-21.13	12.434	.208	-50.45	8.20
Individual Tutoring	Control Group (no tutoring)	35.88*	12.434	.012	6.55	65.20
	Group Tutoring	21.13	12.434	.208	-8.20	50.45

Based on observed means.
The error term is Mean Square(Error) = 6183.855.
*. The mean difference is significant at the .05 level.

Homogeneous Subsets

Post_GREq

Tukey HSD[a,b]

Tutor_Group	N	Subset	
		1	2
Control Group (no tutoring)	80	574.25	
Group Tutoring	80	589.00	589.00
Individual Tutoring	80		610.13
Sig.		.463	.208

Means for groups in homogeneous subsets are displayed.
Based on observed means.
The error term is Mean Square(Error) = 6183.855.
a. Uses Harmonic Mean Sample Size = 80.000.
b. Alpha = .05.

The first table in the output (Between-Subjects Factors) lists each level of the independent variable and the number of subjects in each level. In this case, each of the three levels of the factor Tutor Group has $n = 80$ subjects, for a total of $n = 240$ subjects. The second table (Descriptive Statistics) provides the means and standard deviations for each level of the factor Tutor group, and also the Grand Mean (G = 591.13), which is in the row labeled 'Total'.

The third table (Tests of Between-Subjects Effects) is the results of the ANOVA on the data. In the table on the preceding page, the relevant areas are highlighted in yellow. The 'Source' column lists each source of variance in the data, between subjects (Tutor_Group), within subjects (Error) and Total. The sums of squares, degrees of freedom, and mean square for each source of variance are listed along with the F-Ratio and the *p*-value under Sig. In this case, $p = .016$, which is less than $\alpha = .05$; hence, the ANOVA suggests a significant difference across the levels of the factor Tutor_Group.

The fourth table (Multiple Comparisons) is where each of the three levels of the independent variable are compared to the other levels, and is used to determine which levels are significantly different from each other. In this case, the only two levels that are significantly different from each other are the Control Group (no tutoring) and the individual Tutoring Group. The mean difference in GRE Quantitative Reasoning scores

between these two groups was 35.88 points, which is associated with p = .012. Finally, the table called Homogeneous Subsets is not very relevant for our purposes here and it can be ignored.

CH 18 Homework Questions

1. *Use the following information to answer the questions below.* A researcher studied the relationship between task difficulty and performance. Twenty subjects worked on an identical task, but five subjects were told that the task was of easy, five were told the task moderately difficult five were told that the task highly difficult, and five were given no information. Scores range from 0 to 10, where higher values indicate better task performance. The data follow:

Easy	Moderate	High	No Info
8	6	2	5
7	6	3	4
5	4	2	5
8	3	5	6
7	6	3	5

a. Calculate the mean for each condition and the grand mean.
b. Determine the degrees of freedom total, the degrees of freedom between groups, the degrees of freedom within groups.
c. Calculate the mean within groups, and the mean square between groups.
d. Calculate the F-Ratio.
e. What is the p-value?
f. Assuming α = .05, is the result of the analysis of variance statistically significant? What decisions do we make with the null and alternate hypotheses?
g. Analyze the nature of the relationship between supposed task difficulty and task performance using the Tukey HSD test. That is, calculate Tukey's HSD value and then use that value to determine which means are significantly different.
h. Compute the value of eta-squared and then Cohen's f. Does the observed Cohen's f value represent a small, medium or large effect?
i. Use Cohen's f and G*Power 3, determine the amount of statistical power to correctly reject the null hypothesis.

2. *Use the following information to answer the questions below.* An investigator tested the relationship between drug intake and anxiety. Eighteen subjects who had been diagnosed with anxiety were given a different level of an antianxiety drug (100 mg/day, 200 mg/day, or 300 ms/day). Each subject had the same baseline level of anxiety, and each subject's level of anxiety was re-measured after three months taking the drug. Scores range from 0 to 10, with higher values indicating more anxiety. The data follow:

100 mg/day	200 mg/day	300 mg/day
8	6	2
5	7	1
6	5	0
8	6	2
9	5	2
9	7	2

a. Calculate the mean for each condition and the grand mean.
b. Determine the degrees of freedom total, the degrees of freedom between groups, the degrees of freedom within groups.
c. Calculate the sum of squares total, the sum of squares within groups, and the sum of square between groups.
d. Calculate the mean within groups, and the mean square between groups.

e. Calculate the F-Ratio.
f. What is the p-value?

g. Assuming α = .01, is the result of the analysis of variance statistically significant? What decisions do we make with the null and alternate hypotheses?
h. Analyze the nature of the relationship between supposed task difficulty and task performance using the Tukey HSD test. That is, calculate Tukey's HSD value and then use that value to determine which means are significantly different.
i. Compute the value of eta-squared and then Cohen's f. Does the observed Cohen's f value represent a small, medium or large effect?
j. Using Cohen's f and G*Power 3, determine the amount of statistical power to correctly reject the null hypothesis.

3. For each of the following situations, find the total number of subjects that would be needed to achieve the desired level of Power.

a. k = 3; f = .30; α = .05; Power = .80
b. k = 3; f = .30; α = .05; Power = .95
c. k = 4; f = .20; α = .01; Power = .80
d. k = 4; f = .20; α = .05; Power = .80

4. For each of the following situations, find the amount of Power, based on the parameters given.

a. k = 2; f = .25; α = .05; *n* = 75
b. k = 2; f = .40; α = .01; *n* = 180
c. k = 4; f = .45; α = .05; *n* = 30
d. k = 5; f = .35; α = .01; *n* = 190

5. *Use the following information to answer the questions below.* The bystander effect is the phenomenon where an individual is less likely to help a person when others are around. I examine the bystander effect among Boy Scouts. I randomly select 24 Boy Scouts and assign each to one of four groups. In each group the scout is seated at a table with my assistant, but the scout believes my assistant is another subject. While the scout and my assistant complete paperwork, my assistant falls out of his chair and I record the time (in seconds) it takes the scout to get help. I vary the number of people seated at the table. In Group A only the one scout and my assistant are present (one bystander); in Group B the scout, my assistant, and one other person are present (two bystanders); in Group C the scout, my assistant, and three others are present (four bystanders); and in Group D the scout, my assistant, and seven others are present (eight bystanders). The time, in seconds, it took each scout to get help are below.

Group A (1 Bystander)	**Group B (2 Bystanders)**	**Group C (4 Bystanders)**	**Group D (8 Bystanders)**
7	7	7	14
8	8	8	11
6	12	7	15
5	11	12	16
5	8	11	17
11	8	9	17

a. Calculate the mean for each condition and the grand mean.
b. Determine the degrees of freedom total, the degrees of freedom between groups, the degrees of freedom within groups.
c. Calculate the sum of squares total, the sum of squares within groups, and the sum of square between groups.
d. Calculate the mean within groups, and the mean square between groups.
e. Calculate the F-Ratio.
f. What is the p-value?

g. Assuming α = .01, is the result of the analysis of variance statistically significant? What decisions do we make with the null and alternate hypotheses?
h. Analyze the nature of the relationship between supposed task difficulty and task performance using the Tukey HSD test. That is, calculate Tukey's HSD value and then use that value to determine which means are significantly different.
i. Compute the value of eta-squared and then Cohen's f. Does the observed Cohen's f value represent a small, medium or large effect?
j. Using Cohen's f and G*Power 3, determine the amount of statistical power to correctly reject the null hypothesis.

Chapter 19: Factorial ANOVA

19.1 What are a Factorial Design and Factorial ANOVA?

A **factorial design** includes more than one independent variable and the levels of each independent variable are 'crossed' or 'combined' with the levels of the other independent variables. A simple example of this concept is presented in the picture to the right. Think of the bottle of pills (Pharmaceutical consumption) and the bottles and cans of beer (Alcohol consumption) as two different independent variables that a person may or may not take. Taken alone, each independent variable has its own (independent) influence on the person doing the consuming: both the pills and the alcohol cause drowsiness. But, when the pills and the alcohol are consumed together (independent variables are combined), something new happens: the consumer dies. Hence, when combined the effects of taking pills and drinking alcohol *interact* to produce a new influence on an individual…*death*! This is what a factorial design is all about (without killing of subjects). **Factorial experimental designs** combine the influence of two or more independent variables to examine their combined influence on a dependent variable. Indeed, the purpose of factorial designs to t examine whether there is a combined influence of impendent variables, or not. To analyze factorial designs, we use **factorial ANOVA**. Generally, factorial, between-subjects ANOVA is used to assess the relationship between variables when

1. The dependent variables is *quantitative*
2. Both of the independent variables are *qualitative* in nature
3. Both of the independent variables are manipulated between-subjects
4. Each independent variable has two or more levels
5. The level of each independent variable is combined with levels of the other independent variable

To generalize the example from above, say that I have two independent variables, 'J' and 'K,' and each independent variable has two levels (J_1, J_2, K_1, and K_2). If each level of 'J' is crossed/combined with each level of 'K', I have a factorial design. In this case, because there are two levels of J (J_1 and J_2) and two levels of K (K_1 and K_2) there are 2 (levels of J) x 2 (levels of K) = 4 total combinations/conditions, which can be seen in the table below. This type of a design with two independent variables each with two levels is called **2 by 2 factorial design** and is the simplest factorial design. This type of factorial design will be analyzed using ANOVA in this chapter.

		Variable J	
		J_1	J_2
Variable K	K_1	J_1K_1	J_2K_1
	K_2	J_1K_2	J_2K_2

With a factorial design, several effects *may* be statistically significant in the ANOVA performed on the data. First, there can be a significant **main effect** of each independent variable. A main effect occurs when the mean difference between the levels of an independent variable is statistically significant. For example, a main effect of 'J' would indicate that the difference between the mean of J_1 and the mean of J_2 is statistically significant. In the case of a 2 x 2 factorial design, which has two independent variables, there are two potential main effects: A main effect of J and a main effect of K.

An **interaction** between variables may also be present, and an interaction occurs between the independent variables. An interaction is found when the influence of one independent variable on the dependent variable changes across the levels of another independent variable. States differently, the effect of one independent variable on the dependent variable depends on the levels other independent variables. For example, the difference between J_1 and J_2 may differ across the levels of K. The table below illustrates what an interaction may look like if the difference between J_1 and J_2 changes across levels of K. Specifically, the mean difference between J_1 and J_2 at level K_1 is only a difference of 5; but the difference between J_1 and J_2 at level K_2 is a difference of 40. Thus, the influence of 'J' on the dependent variable is changing across levels of K. This is an interaction.

		Variable J		
		J_1	J_2	Difference
Variable K	K_1	90	85	5
	K_2	50	10	40

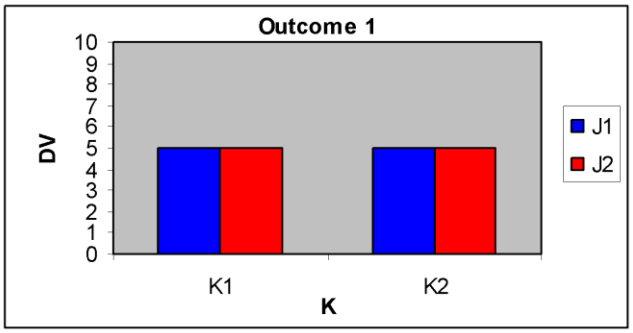
The number of potential interactions depends on the number of independent variables in the factorial design. In the case of a 2 by 2 factorial design, there is the potential for only *one* interaction. If there were three independent variables (e.g., J, K, L) then there are *four* potential interactions: (1) between J and K; (2) between J and L; (3) between K and L; and (4) between J, K, and L. Thus, the number of potential interactions depends on the number of combinations of variables, including combinations of more than two independent variables.

19.2 What do Main Effects and Interactions Look Like?

What do main effects and interactions look like in a 2 x 2 design? That is, what do significant main effects and interactions look like? There are many potential outcome patterns of data, but there are only eight possible outcomes based on the statistical significance of the main effects and an interaction. IN the tables and graphs below, I have displayed hypothetical examples of what these eight outcomes will look like. Take some time to go through these. For each of these outcomes, assume that the measurement scale is 0 - 10. The means for each level of 'J' and for each level of 'K' are also given, as are the differences between J_1 and J_2 to help show any potential interaction.

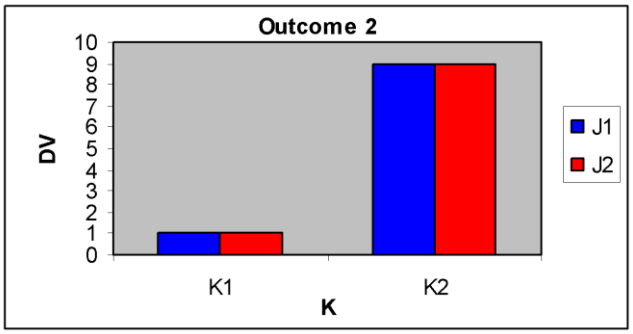
Outcome 1: No Main effects, No Interaction.

		Variable J			
		J_1	J_2	$\overline{K}_k$	Difference
Variable K	K_1	5	5	5	0
	K_2	5	5	5	0
	$\overline{J}_j$	5	5		



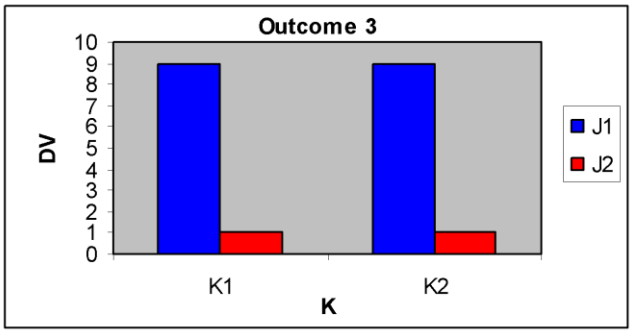
Outcome 2: Main effect of K only.

		Variable J			
		J_1	J_2	$\overline{K}_k$	Difference
Variable K	K_1	1	1	1	0
	K_2	9	9	9	0
	$\overline{J}_j$	5	5		



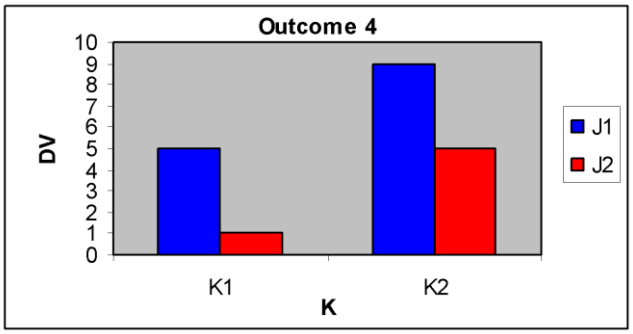
Outcome 3: Main effect of J only

		Variable J			
		J_1	J_2	$\overline{K}_k$	Difference
Variable K	K_1	9	1	5	8
	K_2	9	1	5	8
	$\overline{J}_j$	9	1		



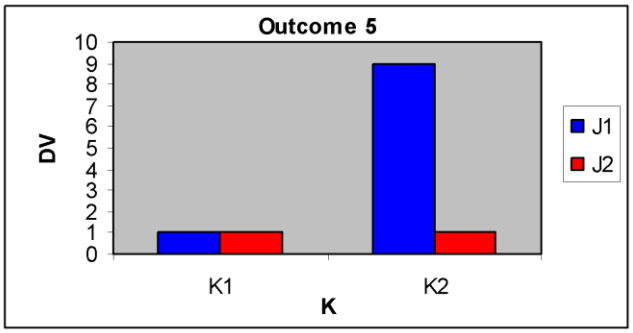
Outcome 4: Main effects of J and K.

		Variable J			
		J_1	J_2	$\overline{K}_k$	Difference
Variable K	K_1	5	1	3	4
	K_2	9	5	7	4
	$\overline{J}_j$	7	3		



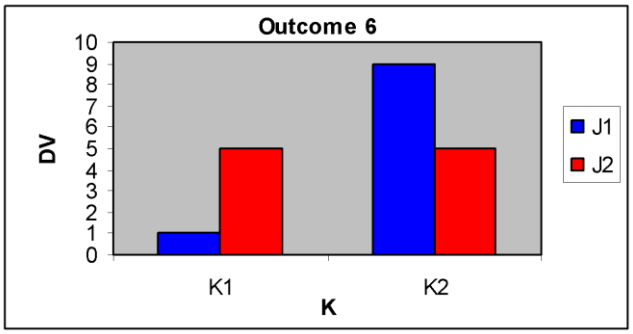
Outcome 5: Main effects of J and K with interaction.

		Variable J			
		J_1	J_2	$\overline{K}_k$	Difference
Variable K	K_1	1	1	1	0
	K_2	9	1	5	8
	$\overline{J}_j$	5	1		



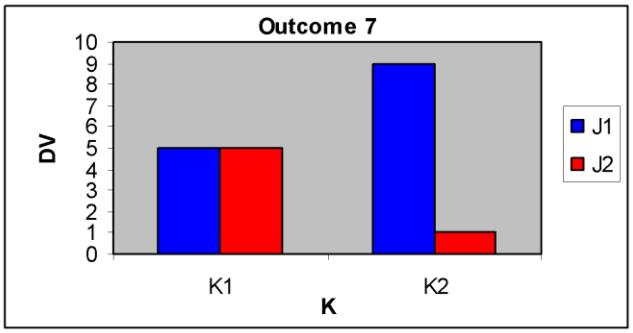
Outcome 6: Main effect of K with interaction.

		Variable J			
		J_1	J_2	$\overline{K}_k$	Difference
Variable K	K_1	1	5	3	-4
	K_2	9	5	7	4
	$\overline{J}_j$	5	5		



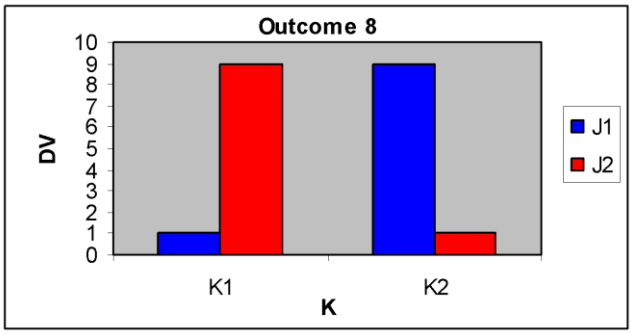
Outcome 7: Main effect of J with interaction.

		Variable J			
		J_1	J_2	$\overline{K}_k$	Difference
Variable K	K_1	5	5	5	0
	K_2	9	1	5	8
	$\overline{J}_j$	7	3		



Outcome 8: Main effects, with interaction.

		Variable J			
		J_1	J_2	$\overline{K}_k$	Difference
Variable K	K_1	1	9	5	-8
	K_2	9	1	5	8
	$\overline{J}_j$	5	5		



19.3 Hypotheses in Factorial Designs

With a factorial design you have a hypothesis for each main effect and each interaction. The null hypothesis for a main effect is that means between the levels of an independent variable will be equal; whereas the alternate hypothesis for a main effect states that the means will not be equal. Thus, for the examples using Variables J and K above, the null and alternate hypotheses:

H_0: $\mu_{J1} = \mu_{J2}$ $\quad\quad$ H_0: $\mu_{K1} = \mu_{K2}$

H_1: $\mu_{J1} \neq \mu_{J2}$ $\quad\quad$ H_1: $\mu_{K1} \neq \mu_{K2}$

For an interaction, the hypotheses are a little trickier. First, you must decide which independent variable is your **focal variable**, that is, the variable you believe will produce a direct change in performance on the dependent variable. The other variable will be your **moderator variable**, which is the variable you believe will influence the relationship between the independent variable and the dependent variable. Say that 'K' is the moderator variable and 'J' is the focal variable. Thus, we would be interested in the difference between J_1 and J_2 at each level of K. If the interaction is not significant, the difference between J_1 and J_2 will be equal at each level of K. If the interaction is significant, the difference between J_1 and J_2 will differ at each level of K. Thus, the null hypothesis for the interaction can be stated as:

$$H_0: (\mu_{J1K1} - \mu_{J2K1}) = (\mu_{J1K2} - \mu_{J2K2})$$

The term before the equal sign is the mean difference between J_1 and J_2 at level K_1, and the term after the equal sign is the difference between J_1 and J_2 at level K_2. The alternate hypothesis is:

$$H_0: (\mu_{J1K1} - \mu_{J2K1}) \neq (\mu_{J1K2} - \mu_{J2K2})$$

Thus, the alternate hypothesis is that the difference between the cell means J_1 and J_2 will not be equal across levels of independent variable K.

19.4 Source Table

In a factorial design, because there is more than one potentially statistically significant effect, we must calculate an F-Ratio for each main effect and also the interaction. For the most part there is little difference in the procedures for conducting a factorial ANOVA from the procedures used to conduct a oneway ANOVA in chapter 19. It is good to create a source table to list each sums of squares, degrees of freedom, and mean square that will be used in your F-Ratios.

Variance Source	SS	df	MS	F
Between Groups	??	??	??	??
Variable 'J'	??	??	??	??
Variable 'K'	??	??	??	??
Interaction (J x K)	??	??	??	??
Within Groups (error)	??	??		
Total Variance	??	??		

You can see that several df's, several SS's, several MS's and several F's need to be found. I will start with the df's and then simply show the formulas for calculating the SS's, which will be used in the next section with a numerical example.

The total degrees of freedom (df_T) are equal to $n_T - 1$, where n_T is the total number of subjects in the data. Between group degrees of freedom (df_B) are equal to (j)(k) – 1, where j is the number of levels of variable J and k is the number of levels of variable K. Degrees of freedom for Variable J (df_J) are equal to j – 1, and degrees of freedom for Variable K (df_K) are equal to k – 1. Degrees of freedom for the interaction (df_{JxK}) are equal to (j – 1)(k – 1). Finally, within-group degrees of freedom (df_W) are equal to n_T – (j)(k). The sums of squares for each source of variance are below.

Sum of **squares total** (**SS_T**):

$$SS_T = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n}(X_{ijk} - G)^2$$

Sum of **squares within-groups** (**SS_W**):

$$SS_W = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n}(X_{ijk} - \overline{X}_{jk})^2$$

Sum of **squares between-groups** (**SS_B**):

$$SS_B = \sum_{jk=1}^{JK} n_{jk}(\overline{X}_{jk} - G)^2$$

Sum of **squares for the main effect of J** (**SS_J**):

$$SS_J = \sum_{j=1}^{J} n_j(\overline{X}_j - G)^2$$

Sum of **squares for the main effect of K** (**SS_K**):

$$SS_K = \sum_{k=1}^{K} n_k(\overline{X}_k - G)^2$$

Sum of **squares for the interaction** (**SS_{JxK}**):

$$SS_{JxK} = \sum_{j=1}^{J} \sum_{k=1}^{K} n_{jk} (\overline{X}_{jk} - \overline{X}_{j.} - \overline{X}_{.k} + G)^2$$

Importantly, note that that $SS_T = SS_B + SS_W$ and that $df_T = df_B + df_W$. Also, note that $SS_B = SS_J + SS_K + SS_{J \times K}$ and that $df_B = df_J + df_K + df_{J \times K}$. Also, the mean square for each source of variance is simply the ratio of the sum of squares for that source of variance divided by the degrees of freedom for that source of variance. Finally, the F-Ratio for each main effect and the interaction is calculated by dividing the mean square for that effect by the mean-square within-subjects, which is always the error term for each F-Ratio.

19.5 Numerical Example of 2 x 2 ANOVA

For this example we'll expand on the ANOVA example from Chapter 18, where the location in which a person tried to recall a list of words, relative to where they studied those words, was manipulated between subjects. A researcher manipulates the location where a person studies and where a person is tested so that the locations are the Same (J_1) or Different (J_2). For example, a person studies a list of words in a classroom and is then tested on their memory for those words when they are in the same classroom, or in a different classroom. We'll call this independent variable the *Location* factor (J). The researcher also manipulates the time of day the material is studied and tested. The time of day can be the *Same* (K_1) or *Different* (K_2). We'll call this the *Time* factor (K). For this example, we'll use an alpha level of $\alpha = .05$ to assess the statistical significance of each main effect and the interaction.

There are two independent variables (Location and Time) each with two levels that are combined to create four conditions: (1) Same-Location/Same-Time (J_1K_1); (2) Same-Location/Different-Time (J_1K_2); (3) Different-Location/Same-Time (J_2K_1); and (4) Different-Location/Different-Time (J_2K_2). Because we know from chapter 19 the location where a person studied influenced memory, we will assume the focal variable is Location and the moderator variable is Time. The table below lists the memory scores (number of words correctly recalled) and the summary statistics for the n = 5 subjects in each group:

		Location (J)		
		Same (J_1)	**Different (J_2)**	**Totals**
Time (K)	**Same (K_1)**	9 7 10 8 6 $\sum X_{J1K1} = 40$ $M_{J1K1} = 8$	5 7 3 4 6 $\sum X_{J2K1} = 25$ $M_{J1K1} = 5$	$n_{K1} = 10$ $\sum X_{K1} = 65$ $M_{K1} = 6.5$
	Different (K_2)	6 7 9 7 6 $\sum X_{J1K2} = 35$ $M_{J1K1} = 7$	2 4 0 2 2 $\sum X_{J2K2} = 10$ $M_{J1K1} = 2$	$n_{K2} = 10$ $\sum X_{K2} = 45$ $M_{K2} = 4.5$
	Totals	$n_{J1} = 10$ $\sum X_{J1} = 75$ $M_{J1} = 7.5$	$n_{J2} = 10$ $\sum X_{J2} = 35$ $M_{J2} = 3.5$	$n_T = 20$ $\sum X = 110$ $G = 5.5$

Calculate sum of squares total (SS_T), by subtracting the grand mean from each score, squaring that difference, and adding the squared differences, such that the sum of squares is calculated over all subjects. The table below shows each subject's score in the second column (the J x K group each subject came from is listed in the first column). The grand mean is then subtracted from each score in the third column, then that difference is squared in the fourth column. The summation of the squared differences is SS_T at the bottom ($SS_T = 139$). This is the total variation across all n = 20 subjects across all four groups.

Sum of Squares Total (SS_T)

Group	X_{ijk}	$(X_{ijk} - G)$	$(X_{ijk} - G)^2$
J_1K_1	9	9 - 5.5 = 3.5	12.25
J_1K_1	7	7 - 5.5 =1.5	2.25
J_1K_1	10	10 - 5.5 = 4.5	20.25
J_1K_1	8	8 - 5.5 = 2.5	6.25
J_1K_1	6	6 - 5.5 = .5	0.25
J_1K_2	6	6 - 5.5 = .5	0.25
J_1K_2	7	7 - 5.5 = 1.5	2.25
J_1K_2	9	9 - 5.5 = 3.5	12.25
J_1K_2	7	7 - 5.5 = 1.5	2.25
J_1K_2	6	6 - 5.5 = .5	0.25
J_2K_1	5	5 - 5.5 = -.5	0.25
J_2K_1	7	7 - 5.5 = 1.5	2.25
J_2K_1	3	3 - 5.5 = -2.5	6.25
J_2K_1	4	4 - 5.5 = -1.5	2.25
J_2K_1	6	6 - 5.5 = .5	0.25
J_2K_2	2	2 - 5.5 = -3.5	12.25
J_2K_2	4	4 - 5.5 = -1.5	2.25
J_2K_2	0	0 - 5.5 = -5.5	30.25
J_2K_2	2	2 - 5.5 = -3.5	12.25
J_2K_2	2	2 - 5.5 =-3.5	12.25
			$SS_T = 139$

Next calculate sum of squares within-groups (SS_W). The formula (see below) has you find the difference between an individual subject's score (X_{ijk}) and the mean of that subject's group, do this for each subject's score, and then square each difference. Finally add the squared differences to get SS_W. In the table to the right, the second column lists each subject's score and the third column lists the mean of the group for that subject. The fourth Column lists the difference between each subject's score and the mean of that subject's group. The fourth column is the squared differences and at the bottom of that column, SS_W is shown to be 34. This value is the total within group variation and is equal to the "*total error variance*".

Sum of Squares Within Groups (SS_W)

Group	X_{ijk}	$\bar{X}_{jk}$	$(X_{ijk} - \bar{X}_{jk})$	$(X_{ijk} - \bar{X}_{jk})^2$
J_1K_1	9	8	9 – 8 = 1	1
J_1K_1	7	8	7 – 8 = -1	1
J_1K_1	10	8	10 – 8 = 2	4
J_1K_1	8	8	8 – 8 = 0	0
J_1K_1	6	8	6 – 8 = -2	4
J_1K_2	6	7	6 – 7 = -1	1
J_1K_2	7	7	7 – 7 = 0	0
J_1K_2	9	7	9 – 7 = 2	4
J_1K_2	7	7	7 – 7 = 0	0
J_1K_2	6	7	6 – 7 = -1	1
J_2K_1	5	5	5 – 5 = 0	0
J_2K_1	7	5	7 – 5 = 2	4
J_2K_1	3	5	3 – 5 = -2	4
J_2K_1	4	5	4 – 5 = -1	1
J_2K_1	6	5	6 – 5 = 1	1
J_2K_2	2	2	2 – 2 = 0	0
J_2K_2	4	2	4 – 2 = 2	4
J_2K_2	0	2	0 – 2 = -2	4
J_2K_2	2	2	2 – 2 = 0	0
J_2K_2	2	2	2 – 2 = 0	0
				$SS_W = 34$

Next calculate the total between-group sum of squares (SS_B) and the sum of squares for the main effect of each independent variable (SS_J and SS_K), and the sum of squares for the interaction effect ($SS_{J \times K}$). The formula for SS_B below has you find the difference between each group (cell) mean and the grand mean and then square the differences. Then, you multiply the squared difference by the number of subjects in the group (cell), which is n = 5 in this example. Finally, add the products to give you SS_B.

The table below lists each J x K group mean from the example above. The third column shows the difference between each cell mean and the grand mean, the fourth column shows the square of that difference, and the fifth column shows the product of that squared difference with the sample size of each group. The value at the bottom ($SS_B = 105$), is the total variance between the four groups.

Sum of Squares Between Groups (SS_B)				
Group	$\bar{X}_{JK}$	$(\bar{X}_{JK} - \bar{\bar{X}})$	$(\bar{X}_{JK} - \bar{\bar{X}})^2$	$n_{JK}(\bar{X}_{JK} - \bar{\bar{X}})^2$
J_1K_1	8	8 – 5.5 =2.5	6.25	5(6.25) = 31.25
J_1K_2	7	7 – 5.5 =1.5	2.25	5(2.25) = 11.25
J_2K_1	5	5 – 5.5 = -0.5	0.25	5(0.25) = 1.25
J_2K_2	2	2 – 5.5 =-3.5	12.25	5(12.25) = 61.25
				SS_B = 105

Now calculate the sum of squares for the main effect of Location (SS_J). This procedure will be identical to that for finding the sum of squares for the main effect of Time (SS_K); the only difference will be that instead of the mean of each level of J (for SS_K), we'll use the mean for each level of K for SS_K. The formula has you find the mean difference between the grand mean and the mean of each level of the independent variable Location, which we have labeled J. The mean of each level of J is shown in the second column of the table below. The third column shows the calculation of the mean difference for each level of 'J'. The fourth column shows the squared mean difference between the grand mean and each level of J. You next multiply the squared difference by the number of subjects associated with that level of 'J', which will be n = 10 for this example. Finally, you add up those products to get SS_J = 80 for this example:

Sum of Squares for Main Effect of J (SS_J)				
Group	$\bar{X}_J$	$(\bar{X}_J - \bar{\bar{X}})$	$(\bar{X}_J - \bar{\bar{X}})^2$	$n_J(\bar{X}_J - \bar{\bar{X}})^2$
J_1	7.5	7.5 – 5.5 = 2	4	10(4) = 40
J_2	3.5	3.5 – 5.5 = -2	4	10(4) = 40
				SS_J = 80

Next, calculate the sum of squares for the Time factor (SS_K). Like finding SS_J, you first find the mean difference between the grand mean and the mean of each level of the independent variable Time (K_1 and K_2), which is shown in the third column in the table below. Next, square the mean difference and then multiply the squared mean difference by the number of subjects in that level of K, which is n = 10 in this example. Finally, add the products to get SS_K = 80:

Sum of Squares for Main Effect of K (SS_K)				
Group	$\bar{X}_K$	$(\bar{X}_K - \bar{\bar{X}})$	$(\bar{X}_K - \bar{\bar{X}})^2$	$n_K(\bar{X}_K - \bar{\bar{X}})^2$
K_1	6.5	6.5 – 5.5 = 1	1	10(1) = 10
K_2	4.5	4.5 – 5.5 = -1	1	10(1) = 10
				SS_K = 20

Finally, calculate the sum of squares for the interaction (SSJ x K). Take each J x K group mean (each cell mean) and subtract the mean for the level of variable J associated with the group. For example, group J_1K_1 has a mean of 8. This group is part of level J_1, which has a mean of 7.5; hence, you subtract that mean (7.5) from 8. Next, subtract the mean for the level of variable K associated with the group. In the case, group J_1K_1 is part of level K_1, which has a mean of 6.5. Next, add the grand mean (5.5) to the value. In the table below, the mean of each J x K group is listed with the mean of each level of J and K that that particular group is part of. The value in the fourth column is the value that is obtained after subtracting the mean of the levels of variable J and variable K that are associated with that particular group, and then adding the grand mean. The fifth column is the squared value of column 4; and the sixth column is the product of column 5 and the sample size associated with the J x K group, which is always be n = 5 in this example. Adding the products in the sixth column gives you $SS_{J \times K}$ = 5 in this example:

Sum of Squares for Interaction ($SS_{J \times K}$)						
Group	$\bar{X}_{JK}$	$\bar{X}_{J.}$	$\bar{X}_{.K}$	$(\bar{X}_{JK} - \bar{X}_{J.} - \bar{X}_{.K} + \bar{\bar{X}})$	$(\bar{X}_{JK} - \bar{X}_{J.} - \bar{X}_{.K} + \bar{\bar{X}})^2$	$n_{JK}(\bar{X}_{JK} - \bar{X}_{J.} - \bar{X}_{.K} + \bar{\bar{X}})^2$
J_1K_1	8	7.5	6.5	8 - 7.5 - 6.5 + 5.5 = -0.5	0.250	5(.25) = 1.250

J_1K_2	7	7.5	4.5	7 - 7.5 - 4.5 + 5.5 = 0.5	0.250	5(.25) = 1.250
J_2K_1	5	3.5	6.5	5 - 3.5 - 6.5 + 5.5 = 0.5	0.250	5(.25) = 1.250
J_2K_2	2	3.5	4.5	2 - 3.5 - 4.5 + 5.5 = -0.5	0.250	5(.25) = 1.250
						$SS_{JxK} = 5$

Remember that $SS_B = SS_J + SS_K + SS_{J \times K}$. Indeed, you can see from the sums of squares that we just calculated that 105 = 80 + 20 + 5. Also remember that $SS_T = SS_B + SS_W$, and indeed you can see that 139 = 105 + 34.

With the sums of squares and degrees of freedom for each source of variance, the next step is to calculate the **mean squares**. Remember, the mean square is simply the sum of squares for a source of variance divided by the degrees of freedom for that course of variance:

$$MS = \frac{SS}{df}$$

The mean square between-groups (MS_B) is not absolutely necessary to calculate, but the mean squares for both independent variables and the interaction term are necessary. However, strictly speaking, you should determine whether there is significant between group variability *before* determining where that significant influence is coming from.

The ANOVA source table below lists the mean square for the main effects of each independent variables (MS_J and MS_K), the interaction ($MS_{J \times K}$), and between groups. Remember, all that you do to calculate the mean squares for each source of variance is divide the sum of squares by its associated degrees of freedom value. The degrees of freedoms are:

Degrees of freedom total: $df_T = n_T - 1 = 20 - 1 = 19$
Degrees of freedom within: $df_W = n_T - jk = 20 - (2)(2) = 20 - 4 = 16$
Degrees of freedom between: $df_B = jk - 1 = (2)(2) - 1 = 4 - 1 = 3$
Degrees of freedom for Location: $df_J = j - 1 = 2 - 1 = 1$
Degrees of freedom for Time: $df_K = k - 1 = 2 - 1 = 1$
Degrees of freedom for Location x Time: $df_{J \times K} = (j - 1)(k - 1) = (2 - 1)(2 - 1) = 1$

From the sums of squares and the mean squares, here are the mean squares:

Variance Source	SS	df	MS	F
Between Groups	105	3	105/3 = 35	??
Location (J)	80	1	80/1 = 80	??
Time (K)	20	1	20/1 = 20	??
Location x Time (J x K)	5	1	5/1 = 5	??
Within Groups (error)	34	16	34/16 = 2.125	
Total Variance	139	19		

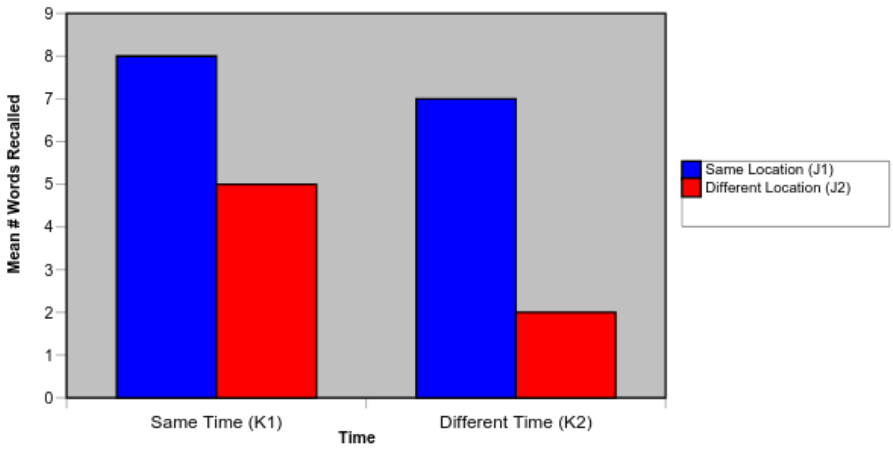
The final step is to calculate the necessary F-Ratios for the main effects and for the interaction. Each F-Ratio is found by dividing the mean square associated with each effect by MS_W, which is the error term for an F-Ratio. We have three effects that can be significant (main effect of Location, main effect of Time, and a Location by Time interaction); thus, we need three F-Value. The F-Ratio for the between-groups source of variance is not absolutely necessary, but I have included it below for completeness. The F-Ratios are listed in the ANOVA summary table below:

Variance Source	SS	df	MS	F
Between Groups	105	3	35	35/2.125 = 16.470
Location (J)	80	1	80	80/2.125 = 37.647
Time (K)	20	1	20	20/2.125 = 9.412
Location x Time (J x K)	5	1	5	5/2.125 = 2.353
Within Groups (error)	34	16	2.125	
Total Variance	139	19		

To determine whether each main effect and the interaction is statistically significant, we must look up a separate p-value for each F-Ratio in Table 3 in Appendix A. Because $df_B = 1$, we'll use Table 3-A, and because the F-Ratio for the Location factor is greater than 10, we'll use F = 10 to assess the statistical significance of the main effect of location. (You do not assess the statistical significance of the overall between groups variance.) The table below lists the p-values for each main effect and the interaction:

Variance Source	SS	df	MS	F	*p*
Location (J)	80	1	80	37.647	.0060
Time (K)	20	1	20	9.412	.0074
Location x Time (J x K)	5	1	5	2.353	.1448
Within Groups (error)	34	16	2.125		---
Total Variance	139	19			---

The p-value for each main effect is less than the selected alpha level (.05), so both main effects are statistically significant. However, the p-value for the interaction (.1448) is greater than the alpha level, so the interaction is not statistically significant. What does it mean? A significant main effect suggests there is a significant mean difference between at least two levels for the independent variable. In the present example, because each variable has only two levels the significant main effect suggests that the mean difference between the levels of each independent variable is significant. Thus, for the variable 'Location' the mean of the 'Same' condition (7.5) and the 'Different' condition (3.5) significantly differ; and for the variable 'Time' the mean of the 'Same' condition (6.5) and the 'Different' condition (4.5) significantly differ. You can see these main effects (and the lack of an interaction), in the graph of the data below:



19.6 Effect Size and Explained Variance

You can also find the effect size for each main effect and the interaction. Each of these values, below, represents the proportion of variance explained in the relationship between the independent variable, or both independent variables in the interaction, with the dependent variable.

$$\eta^2_J = \frac{SS_J}{SS_T} = \frac{80}{139} = 0.576 \quad \eta^2_K = \frac{SS_K}{SS_T} = \frac{20}{139} = 0.144 \quad \eta^2_{JxK} = \frac{SS_{JxK}}{SS_T} = \frac{5}{139} = 0.036$$

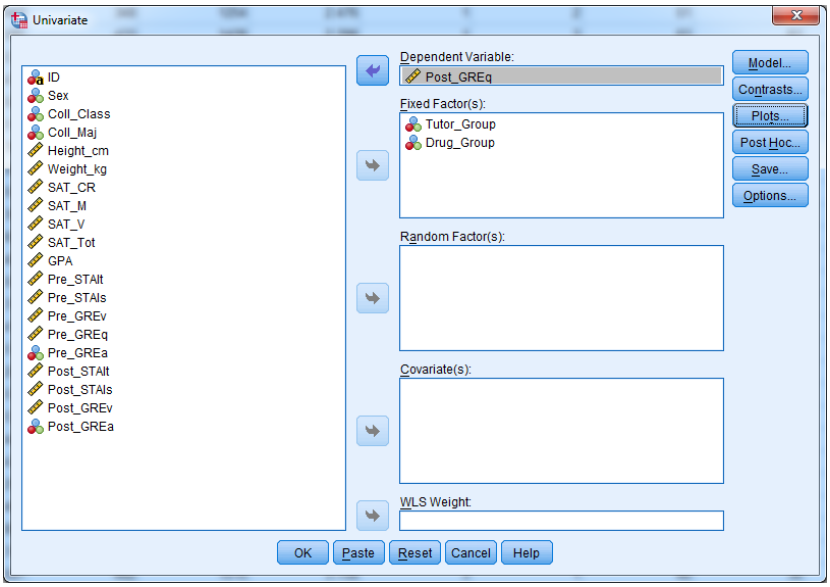
19.7 Factorial ANOVA in SPSS

The following uses the GRE Therapy Data file. Recall, this data file is based on a hypothetical study examining the influences of a study-aid drug and types of tutoring on performance on the Graduate Record

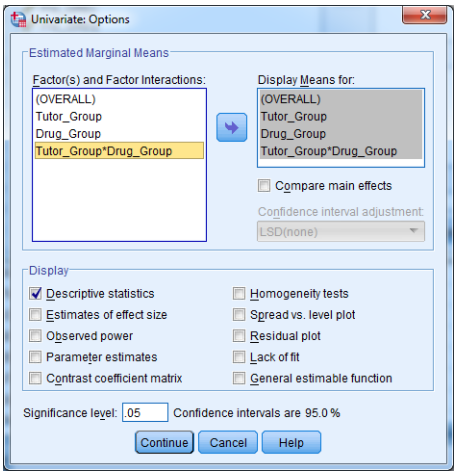
Examinations (GREs). The data file includes a number of dependent variables and as well as the two independent variables (Drug_Group and Tutor_Group). In particular, GRE scores were obtained *prior* to introducing the independent variables (Pre_GREv, Pre_GREq, Pre_GREa) and *after* introducing the independent variables (Post_GREv, Post_GREq, Post_GREa).

This example will build on the ANOVA from Chapter 18 that examined whether the post-test GRE Quantitative Scores (Post_GREq) differed across the levels of the independent variable Tutor_Group. In this data set, there are three levels of the variable Tutor Group: No Tutoring, Group Tutoring, and Individual Tutoring. IN this example, we'll examine the effects of both Tutror_Group and Drug_Group, which has four levels (Control Group, Placebo Group, 100 mg/day, 200 mg/day), on GRE Quantitative Scores.

To request SPSS perform an factorial ANOVA, from the Analyze menu, select General Linear Model, and the Univariate. In the window that opens, you must declare the dependent variable (Post_GREq) in the area under Dependent Variables, and then list both of the independent variables in the area under Fixed Factors:



Click the 'Options…' button and be sure to check off Descriptive Statistics. You'll also want to highlight each of the Factors and Factor Interactions on the left and click the arrow button to move them to the right (see figure at right). Click Continue and then click OK in the main window to have SPSS run the ANOVA; the output should look like that below. Most of the tables in the output were explained in the output produced in Chapter 18. Importantly, you should notice that in the Tests of Between Subjects Effects Table, the main effect of tutor group is statistically significant (as it was in Chapter 18), but not the main effect of Drug Group or interaction.



Univariate Analysis of Variance

Between-Subjects Factors

		Value Label	N
Tutor_Group	1	Control Group (no tutoring)	80
	2	Group Tutoring	80
	3	Individual Tutoring	80
Drug_Group	1	Control Group (no drug)	60
	2	Placebo Group	60

3	100 mg/day Group	60
4	200 mg/day Group	60

Descriptive Statistics

Dependent Variable: Post_GREq

Tutor_Group	Drug_Group	Mean	Std. Deviation	N
Control Group (no tutoring)	Control Group (no drug)	560.00	78.940	20
	Placebo Group	589.00	71.517	20
	100 mg/day Group	575.00	82.430	20
	200 mg/day Group	573.00	75.818	20
	Total	574.25	76.502	80
Group Tutoring	Control Group (no drug)	609.50	76.810	20
	Placebo Group	578.00	66.221	20
	100 mg/day Group	597.00	75.888	20
	200 mg/day Group	571.50	77.682	20
	Total	589.00	74.436	80
Individual Tutoring	Control Group (no drug)	596.50	88.334	20
	Placebo Group	607.00	93.533	20
	100 mg/day Group	597.50	75.105	20
	200 mg/day Group	639.50	79.305	20
	Total	610.13	84.606	80
Total	Control Group (no drug)	588.67	82.861	60
	Placebo Group	591.33	77.601	60
	100 mg/day Group	589.83	77.273	60
	200 mg/day Group	594.67	82.718	60
	Total	591.13	79.685	240

Tests of Between-Subjects Effects

Dependent Variable: Post_GREq

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	103061.250[a]	11	9369.205	1.510	.129
Intercept	83862903.750	1	83862903.750	13517.334	.000
Tutor_Group	52022.500	2	26011.250	4.193	.016
Drug_Group	1217.917	3	405.972	.065	.978
Tutor_Group * Drug_Group	49820.833	6	8303.472	1.338	.241
Error	1414535.000	228	6204.101		
Total	85380500.000	240			
Corrected Total	1517596.250	239			

a. R Squared = .068 (Adjusted R Squared = .023)

Estimated Marginal Means

1. Grand Mean

Dependent Variable: Post_GREq

Mean	Std. Error	95% Confidence Interval	
		Lower Bound	Upper Bound
591.125	5.084	581.107	601.143

2. Tutor_Group

Dependent Variable: Post_GREq

Tutor_Group	Mean	Std. Error	95% Confidence Interval

			Lower Bound	Upper Bound
Control Group (no tutoring)	574.250	8.806	556.898	591.602
Group Tutoring	589.000	8.806	571.648	606.352
Individual Tutoring	610.125	8.806	592.773	627.477

3. Drug_Group

Dependent Variable: Post_GREq

Drug_Group	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Control Group (no drug)	588.667	10.169	568.630	608.703
Placebo Group	591.333	10.169	571.297	611.370
100 mg/day Group	589.833	10.169	569.797	609.870
200 mg/day Group	594.667	10.169	574.630	614.703

4. Tutor_Group * Drug_Group

Dependent Variable: Post_GREq

Tutor_Group	Drug_Group	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Control Group (no tutoring)	Control Group (no drug)	560.000	17.613	525.296	594.704
	Placebo Group	589.000	17.613	554.296	623.704
	100 mg/day Group	575.000	17.613	540.296	609.704
	200 mg/day Group	573.000	17.613	538.296	607.704
Group Tutoring	Control Group (no drug)	609.500	17.613	574.796	644.204
	Placebo Group	578.000	17.613	543.296	612.704
	100 mg/day Group	597.000	17.613	562.296	631.704
	200 mg/day Group	571.500	17.613	536.796	606.204
Individual Tutoring	Control Group (no drug)	596.500	17.613	561.796	631.204
	Placebo Group	607.000	17.613	572.296	641.704
	100 mg/day Group	597.500	17.613	562.796	632.204
	200 mg/day Group	639.500	17.613	604.796	674.204

CH 19 Homework Questions

1. What is a factorial experimental design? What result can be observed in a factorial design that cannot be observed in a single-factor experimental design?

2. Differentiate between a main effect and an interaction.

3. How many groups are in a 3 x 3 factorial design? How many independent variables?

4. How many groups are in a 3 x 2 x 5 factorial design? How many independent variables?

5. How many groups are in a 2 x 2 x 2 x 2 factorial design? How many independent variables?

6. For each of the following sets of population means, indicate whether there is a main effect of factor A, a main effect of factor B, and/or an interaction effect:

a.

	B_1	B_2
A_1	4.00	5.00
A_2	4.00	5.00

Main Effect of A:
Main Effect of B:
Interaction Effect:

b.

	B_1	B_2
A_1	6.00	6.00
A_2	4.00	4.00
A_3	7.00	7.00

Main Effect of A:
Main Effect of B:
Interaction Effect:

c.

	B_1	B_2
A_1	5.00	2.00
A_2	2.00	5.00

Main Effect of A:
Main Effect of B:
Interaction Effect:

d.

	B_1	B_2
A_1	3.00	3.00
A_2	8.00	8.00

Main Effect of A:
Main Effect of B:
Interaction Effect:

Use the following information to answer #7 – 9: For each of the following sets of population means, indicate whether there is a main effect of factor A, a main effect of factor B, and/or an interaction effect.

7. Dr. Kruger has developed an herbal supplement that when taken before an exam should reduce anxiety and improve exam performance. Dr. Kruger pretests 100 students on an test that is designed to measure 'test anxiety' (Score Range 10-90). He then randomly assigns half of the students to an "Experimental" group that receives a daily dosage of the supplement and the other half of the students to a "Control" group that does not receive the supplement. After one month Dr. Kruger retests all the students on a the anxiety test (posttest). The means are:

Group	Pretest	Posttest
Experimental	75	40
Control	70	65

a. Is there a main effect of Pretest-Posttest?

b. Is there a main effect of Group?
c. Is there an interaction?

8. Dr. Keenan believes that consuming LSD and listening to Pink Floyd helps theoretical physicists understand the concept "matter is simply energy condensed to a slow steady vibration." To examine this issue Dr. Keenan samples 200 theoretical physicist and randomly assigns each physicist to one of four groups created by combining two variables. These four groups differ in whether the theoretical physicists (a) consume LSD or not, and (b) listen to Pink Floyd or not. In each group, Dr. Keenan explains his concept and has each theoretical physicist take a test (Score Range: 0-50) that assesses how well they understand the concept "matter is simply energy condensed to a slow steady vibration." The means of each group are:

LSD	Listening to Pink Floyd	Not Listening to Pink Floyd
Consuming	48	38
Not Consuming	35	10

a. Is there a main effect of Listening to Pink Floyd?
b. Is there a main effect of Consuming LSD?
c. Is there an interaction?

9. Dr. Feely believes that cognitive behavioral therapy works just as good at alleviating an individual's depression as drug therapy. To examine this, he randomly samples individuals that have just been diagnosed with moderate depression and have the same score on the Beck Depression Inventory (Score Range: 0-63), but have not started receiving any treatment. He then randomly assigns each individual to one of four groups that differ in whether the individuals (a) receive cognitive behavioral therapy or standard therapy, and (b) receive an antidepressant or not. After two months, he re-measures each individual's level of depression using the Beck Depression Inventory. The mean Beck Depression Inventory scores are:

Antidepressant	Cognitive Behavioral Therapy	Standard Therapy
Taking	10	30
Not Taking	30	50

a. Is there a main effect of Cognitive Behavioral Therapy vs. Standard Therapy?
b. Is there a main effect of Taking the Antidepressant vs. Not Taking the Antidepressant?
c. Is there an interaction?

10. Below are several incomplete ANOVA summary tables. Fill in the missing terms based on the information given. Assume n is equal for each group:

a. Source of Variance	SS	df	MS	F
Between-Groups	60		20	
Main Effect of A	3			
Main Effect of B			7	
Interaction (A x B)			50	
Within Groups	144	36		--
Total			--	--

b. Source of Variance	SS	df	MS	F
Between-Groups		3	136.667	
Main Effect of A	200		200	
Main Effect of B				1
Interaction (A x B)	200			
Within Groups			10	--

Total	1370	99	--	--

c. Source of Variance	SS	df	MS	F
Between-Groups	200	3		
Main Effect of A				
Main Effect of B	50			
Interaction (A x B)	25			
Within Groups			10	--
Total		47	--	--

11. Below are several incomplete ANOVA summary tables. Fill in the missing terms based on the information given. Assume n is equal for each group:

a. Source of Variance	SS	df	MS	F
Between			39	3.9
Main Effect of X	100	1		
Main Effect of Y			2	
Interaction (X x Y)		1		1.5
Within Groups	360	36	10	---
Total			---	---

b. Source of Variance	SS	df	MS	F
Between	151			
Variable X				0.04
Variable Y	50			
Interaction				4
Within		76	25	---
Total	2051	79	---	---

12. Determine the p-value for the F-Ratios for each main effect and the interaction in #10 a – c.

a. Main Effect of A: Main Effect of B: Interaction:
b. Main Effect of A: Main Effect of B: Interaction:
c. Main Effect of A: Main Effect of B: Interaction:

13. Determine the p-value for the F-Ratios for each main effect and the interaction in #11 a – b.

a. Main Effect of X: Main Effect of Y: Interaction:
b. Main Effect of X: Main Effect of Y: Interaction:

14. *Use the following to answer the questions that follow:* An investigator tested the relationship between perceived task difficulty and time limits on performance. Twenty students worked on the same verbal analogy task, but half the students were told that the task was of low difficulty, while the other half were told that that the task was of high difficulty. Half of the students in each of these two groups were given five minutes to complete the task (time limit), and the other half of the students in each difficulty level group were given unlimited time (no limit). Thus, there were four independent groups. Scores on the task could range from 0 to 10, with higher values indicating better task performance. The data are below:

		Task Difficulty (J)	
		Low (J_1)	High (J_2)
Time Limit (K)	No Limit (K_1)	9	6
		8	7
		10	5
		7	8
		6	4
	Limit (K_2)	7	2
		6	3
		4	4
		8	3
		5	3

a. Calculate the mean for each of the four groups, then calculate the mean for each level of each independent variable, and then calculate the grand mean.
b. Calculate each degrees of freedom value.
c. Calculate the sum of squares between groups, the sum of squares within groups and the sum of squares total. Then calculate the sum of squares for each main effect and for the interaction.
d. Calculate each mean square.
e. Calculate each F-Ratio.
f. Determine the p-value for each main effect and the interaction.
g. Based on the p-values and assuming α = .05, what decisions should be made with respect to each main effect and the interaction? What should the investigator conclude with respect to the relationship between task difficulty, time limits, and task performance?

15. *Use the following to answer the questions that follow:* Scores (0 - 10) were obtained from each of four groups that were created by combining the levels of two independent variables. The data are below:

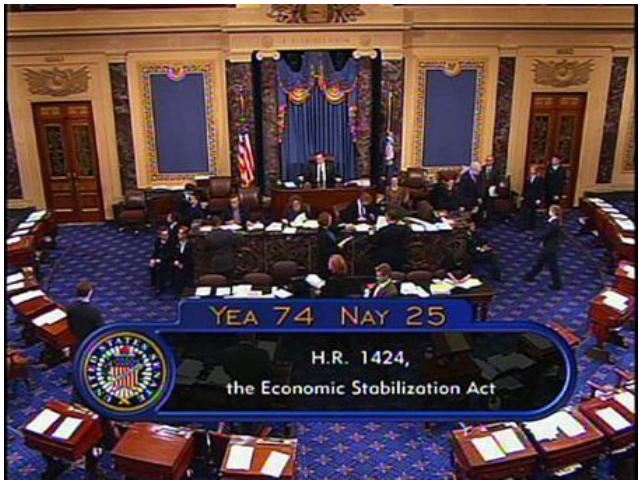
		Variable J	
		J_1	J_2
Variable K	K_1	9	1
		8	3
		10	2
	K_2	3	9
		2	8
		4	7

a. Calculate the mean for each of the four J x K groups. Calculate the mean for each level of each independent variable, and then calculate the grand mean.
b. Calculate each degrees of freedom value.
c. Calculate the sum of squares between groups, the sum of squares within groups and the sum of squares total. Then calculate the sum of squares for each main effect and for the interaction.
d. Calculate each mean square.
e. Calculate each F-Ratio.
f. Determine the p-value for each main effect and the interaction.

Based on each p-value and assuming α = .05, what decision should be made with respect to each main effect and the interaction?

Chapter 20: Chi-Square Analyses

20.1 When, Why, Where is Chi Square Used?



Chi-square (χ^2) analyses are used to examine whether observed data fits with expected data. More specifically, does an observed frequency distribution 'fit' with an expected frequency distribution. As a simple example, look at the picture to the right. It is a snapshot of the television network C-Span at about the end of a US Senate vote on the Economic Stabilization act. Assuming 100 US Senators can vote and assuming they vote 'Yea' or Nay on the Act: how many Senators would vote Yea and vote Nay, *by chance*? If Senators voted Yea and Nay randomly, you would expect, by chance, 50 Yeas and 50 Nays. This is the expected frequency distribution of Yea's and Nay's. Clearly though, there is a difference in the actual (observed) number of Yeas (74) and Nays (25), in the picture above. What a chi-square analysis does is determine whether the observed frequency distribution is significantly different from the expected frequency distribution.

As a simpler, and non-political example: by chance, you would expect that after flipping a coin 10 times, you will observe five heads and five tails. What if you observed nine heads and one tail, would this outcome differ from what is expected by chance? Does this mean that the coin is biased? The chi-square test can help answer this, that is, whether the observed frequencies of heads and tails in the coin-flipping example differ from the frequencies of heads and tails that would be expected to be observed by chance.

In the behavioral sciences, chi-square analyses are often used to evaluate relationships between independent variables that are on a nominal scale and when the collected data are frequencies. Thus, chi-square analyses can be used to evaluate whether frequencies that were observed in a study differ from frequencies that one expects to obtain in a study. For example, a chi-square test can be used to assess whether the numbers of freshmen, sophomores, juniors, and seniors that are enrolled in a course differ from the numbers of each college class that you would expect to be taking the class. For example, in my spring 2011 Cognitive Psychology class, I had 4 freshmen, 15 sophomores, 7 juniors and 8 seniors enrolled. Is this distribution of college classes different from what I should expect? Why are there not equal numbers of students from each college class? A chi-square test can be used to help evaluate this. If I find that the observed frequencies differ from what is expected, I might conclude that there is something about this course that influences students of a certain college class to take the course. Generally, chi-square is used to assess the relationship between variables when

1. The variable, or variables, are *qualitative* in nature (nominal scale)
2. If there are two variables, the same subjects have been measured on each variable
3. The observations on the variable, or variables, are between subjects
4. Each level of each independent variable is combined with each level of the second independent variable

20.2 Assumptions of the Null Hypothesis

Under the null hypothesis, one would assume that a set of **observed frequencies** (**f_o**) will be equal to the **expected frequencies** (**f_e**). That is, if I expect to observe five freshmen in a class, then I should observe five freshmen in that class. Thus, the null hypothesis predicts no difference between the observed frequencies and the expected frequencies, such that the expected chi-square value of under the null hypothesis is zero. Hence, the null hypothesis is:

$$H_0: f_0 = f_e$$

The alternate hypothesis predicts that the observed and the expected frequencies will not be equal:

$$H_1: f_0 \neq f_e$$

There is no such thing as directionality of the alternate hypothesis for chi-square analyses.

20.3 Chi Square with One Independent Variable

Chi-square makes several assumptions. First, chi-square tests assume the dependent variable that is being analyzed is frequency; hence, numbers of observed cases in a given category. Second the levels of each independent variable come from a nominal scale. In the Cognitive Psychology example above, the frequencies of freshmen, sophomores, juniors, and seniors represent groups that differ along a nominal variable (*college class*). Third, levels of each independent variable must be *mutually exclusive*; that is, a person, or whatever is being tallied, can belong to only one level for each independent variable.

The example below deal with a chi-square analysis applied to a single independent variable. Note, the independent variables in each example in this chapter are not manipulated; rather, the independent variable represents groups that differ along some naturally occurring dimension. We will use the chi-square test to determine whether the frequencies observed across the categories of one variable differ from what are expected. Because there is only one variable, this type of a design is a *oneway design*; thus, the test is a **oneway chi-square**.

Say I am interested in the distribution of college majors in my Sensation & Perception class. I examine my class' roster to determine the major for each of the n = 39 students in my class. I find that there are 17 psychology majors, 3 counseling and human service majors, 3 neuroscience majors, 6 occupational therapy majors, 4 communications majors, and 6 students falling into some 'other' major (history, English, etc.). Thus, I classified the students into six mutually exclusive groups based on major. Hence, I have one independent variable (College Major), with six levels. The observed frequencies for each major are:

Major					
Psychology	Counseling & Human Services	Neuroscience	Occupational Therapy	Communications	Other
$f_o = 17$	$f_o = 3$	$f_o = 3$	$f_o = 6$	$f_o = 4$	$f_o = 6$

Remember, chi-square analyses will compare a distribution of observed frequencies (f_o) to a distribution of expected frequencies (f_e). The table above lists the observed frequencies, so the expected frequencies need to be determined. I have n = 39 students in this class and there are six different majors. If choice of college major did not determine whether a student took this course, I would expect there equal numbers of students in each of the six observed majors. Because there are k = 6 majors ('k' is the number levels of the independent variable), I would expect there to be 39/6 = 6.5 students in each college major that are in this course. The formula for figuring out the expected frequencies for each group in a one-way chi-square can be expressed as:

$$f_e = \frac{n}{k}$$

Where n_T is the total number of subjects, and 'k' is the total number of levels of the independent variable. In a one-way chi-square, f_e will be the same for each group. The table below lists the observed and expected frequencies for each group:

Major					
Psychology	Counseling & Human Services	Neuroscience	Occupational Therapy	Communications	Other
$f_o = 17$	$f_o = 3$	$f_o = 3$	$f_o = 6$	$f_o = 4$	$f_o = 6$
$f_e = 6.5$	$f_e = 6.5$	$f_e = 6.5$	$f_e = 6.5$	$f_e = 6.5$	$f_e = 6.5$

Once the distribution of expected frequencies has been determined, a chi-square analysis can be conducted. The formula for the chi-square analysis compares each observed frequency to each expected frequency for each level of the independent variable. Remember, if there is no difference between the expected and observed frequencies, then the value of the chi-square should be equal to 0. The the farther the chi-square value is from zero the more likely there is a significant difference between the observed and the expected frequencies. The formula for the chi-square test is:

$$\chi^2 = \sum\left(\frac{(f_O - f_e)^2}{f_e}\right)$$

First, you first subtract the expected frequency for a level of the independent variable from the observed frequency for that level. You then square the difference and then divide the squared difference by the expected frequency, before finally adding the quotients. This is done in the table below:

Group	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
Psychology	17	6.5	10.5	110.25	16.962
Counseling & Human Services	3	6.5	-3.5	12.25	1.885
Neuroscience	3	6.5	-3.5	12.25	1.885
Occupational Therapy	6	6.5	-0.5	0.25	0.038
Communications	4	6.5	-2.5	6.25	0.962
Other	6	6.5	-0.5	0.25	0.038
					$\chi^2 = 21.77$

To assess the statistical significance of the outcome of a chi-square test, you first calculate the degrees of freedom, which are equal to df = k – 1, where k is the number of levels of the independent variable. In this case there are six categories, so df = 6 – 1 = 5. Assume we select an alpha-level of α = .05. Probabilities under the chi-square distribution are presented in Table 5 in Appendix A, a portion of which appears on the next page. Table 5 is set up similarly to the t-tables (Table 2), where the test statistic values are listed in the leftmost column and the degree of freedom values are listed in the column headings.

In Table 5 locate the column associated with df = 5 in the column headings (highlighted in yellow) and scroll down that column until you locate the row associated with the obtained chi-square value (21.77) or the closest chi-square value that is less than the observed value (21.50, highlighted in yellow). IN this example, the p-value associated with $\chi^2 = 21.77$ is $p = .0007$, which is less than the selected alpha level of .05; hence, the null hypothesis can be rejected and the alternate hypothesis accepted. I conclude there is a significant difference between the distribution of observed frequencies from the distribution of expected frequencies of college majors in my Sensation & Perception class. This means is that there may be some relationship between this course and the types students, based on college major, who take the course. Specifically, certain majors may be more likely to take this particular course.

	Degrees of Freedom (df)																					
χ^2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30
20.00	0.0000	0.0000	0.0002	0.0005	0.0012	0.0028	0.0056	0.0103	0.0179	0.0293	0.0453	0.0671	0.0952	0.1301	0.1719	0.2202	0.2742	0.3328	0.3946	0.4579	0.7468	0.9165
20.50	0.0000	0.0000	0.0001	0.0004	0.0010	0.0023	0.0046	0.0086	0.0151	0.0249	0.0389	0.0582	0.0834	0.1151	0.1536	0.1985	0.2495	0.3054	0.3651	0.4271	0.7201	0.9029
21.00	0.0000	0.0000	0.0001	0.0003	0.0008	0.0018	0.0038	0.0071	0.0127	0.0211	0.0334	0.0504	0.0729	0.1016	0.1368	0.1785	0.2263	0.2794	0.3368	0.3971	0.6926	0.8879
21.50	0.0000	0.0000	0.0001	0.0003	0.0007	0.0015	0.0031	0.0059	0.0106	0.0179	0.0285	0.0435	0.0636	0.0895	0.1216	0.1601	0.2047	0.2549	0.3098	0.3682	0.6644	0.8716
22.00	0.0000	0.0000	0.0001	0.0002	0.0005	0.0012	0.0025	0.0049	0.0089	0.0151	0.0244	0.0375	0.0554	0.0786	0.1078	0.1432	0.1847	0.2320	0.2843	0.3405	0.6357	0.8540
22.50	0.0000	0.0000	0.0001	0.0002	0.0004	0.0010	0.0021	0.0041	0.0074	0.0128	0.0208	0.0323	0.0481	0.0689	0.0953	0.1278	0.1662	0.2105	0.2601	0.3140	0.6067	0.8352
23.00	0.0000	0.0000	0.0000	0.0001	0.0003	0.0008	0.0017	0.0034	0.0062	0.0107	0.0177	0.0277	0.0417	0.0603	0.0841	0.1137	0.1493	0.1906	0.2373	0.2888	0.5776	0.8153
23.50	0.0000	0.0000	0.0000	0.0001	0.0003	0.0006	0.0014	0.0028	0.0052	0.0090	0.0150	0.0238	0.0361	0.0526	0.0741	0.1010	0.1337	0.1721	0.2160	0.2649	0.5484	0.7942
24.00	0.0000	0.0000	0.0000	0.0001	0.0002	0.0005	0.0011	0.0023	0.0043	0.0076	0.0127	0.0203	0.0311	0.0458	0.0651	0.0895	0.1194	0.1550	0.1962	0.2424	0.5194	0.7720
24.50	0.0000	0.0000	0.0000	0.0001	0.0002	0.0004	0.0009	0.0019	0.0036	0.0064	0.0108	0.0174	0.0268	0.0398	0.0571	0.0791	0.1065	0.1393	0.1777	0.2212	0.4907	0.7489
25.00	0.0000	0.0000	0.0000	0.0001	0.0001	0.0003	0.0008	0.0016	0.0030	0.0053	0.0091	0.0148	0.0231	0.0346	0.0499	0.0698	0.0947	0.1249	0.1605	0.2014	0.4624	0.7250

One thing you will notice is there are no means to compare across levels of the independent variable, which means chi-square is a **non-parametric statistic**. How does one perform a *post-hoc analysis* once the chi-square analysis has established a significant difference between the observed and expected frequencies? There are a number of post-hoc analyses that can be performed, but they are beyond the scope of this chapter. Instead, I will describe how you can go about interpreting the results of the chi-square analysis.

You can describe the difference between the observed frequencies and the expected frequencies by examining the $(f_o - f_e)^2/f_e$ values and the $f_o - f_e$ values for each level of the independent variable in the table above. Specifically, because the $(f_o - f_e)^2/f_e$ values are summed to give you the chi-square statistic, larger $(f_o - f_e)^2/f_e$ values will contribute more to the rejection of the null hypothesis. What you do is look for the largest values in the $(f_o - f_e)^2/f_e$ column and then use the corresponding $f_o - f_e$ values to determine how that level of the independent variable contributed to the significant relationship between the independent variable and the observed frequencies

The largest $(f_o – f_e)^2/f_e$ value in that column is for Psychology majors (16.962), while the other values are all quite small (Range = 0.038 – 1.885). Examination of the corresponding $f_o – f_e$ value indicates that psychology majors are *more* likely to take this class than expected ($f_o – f_e$ = 10.5). Although much smaller, the $(f_o – f_e)^2/f_e$ value for Neuroscience and Counseling & Human Services (1.885) also indicate a contribution to the rejection of the null hypothesis. Examination of the corresponding $f_o - f_e$ value indicates that both Neuroscience and Counseling & Human Services majors are *less* likely to take this class than expected ($f_o – f_e$ = -3.5). The point is that you should examining the differences between what is expected and what is observed after finding a significant chi-square statistic, in order to decipher the relationship.

20.4 Goodness of Fit Test

The **chi-square goodness of fit test** is similar to the oneway chi-square test from Section 21.3. The goodness of fit test is used to determine whether the distribution of groups in a sample is the same as the distribution of those groups in a population. That is, we know something about the relative frequencies within groups in a population, and the goodness of fit test is used to determine whether there is a significant deviation of those expected frequencies in a sample. Stated differently, the goodness of fit test is used to determine whether the relative frequencies (proportions) of groups in a population are proportional in a sample.

For example, the relative frequencies (proportions) of females and males living in the United States is about p(Male) = .49 and p(Female) = .51. Say that I need a representative sample of n = 100 subjects to take survey. If I want this sample to represent the United States' population of males and females, I need to have 49 males and 51 females, based on the proportions so males and females in the population (i.e., n_{Males} = .49 x 100 = 49; and $n_{Females}$ = .51 x 100 = 51). What if I end up with 55 females and 45 males? Is this a significant deviation from the numbers that I should expect to find in my sample, based on the population parameters? The chi-square goodness of fit test can answer this.

As another example, it is estimated that the proportion of left-handed people in the United States is p(Left Handed) = .10, and the proportion of right-handed people is p(Right Handed) = .90. We want to know whether these proportions of left-handed and right-handed people hold for all of the individuals who have held the office of President of the United States. That is, over all of the Presidents of the United States, does the proportion of left-handed Presidents and right-handed Presidents deviates from the population proportions?

First, find out how many individuals have been the President of the United States and then how many were left-handed and how many were right-handed. In all, there have been n = 44 Presidents of the United States, including Barack Obama. Of these n = 44 Presidents, there have been eight left-handed Presidents (Barack Obama, Bill Clinton, George Bush Sr., Ronald Reagan, Gerald Ford, Harry Truman, Herbert Hoover, and James Garfield. Each of the other 36 Presidents of the United States was right handed.

Next, determine the expected frequencies (f_e) of left-handed and right-handed Presidents, based on the population proportions of left- and right-handed people and the total number of Presidents (44). The expected frequencies for the goodness of fit test can be found by the following equation:

$$f_e = (p)(n)$$

In this equation, *p* is the relative frequency (proportion) associated with group 'i' in the population. In this example, the *p* value will be the proportion of left-handed people in the United States (.10) and the proportion of right-handed people in the United States (.90). The n is the number of subjects in the sample. In this case the sample size is the n = 44 Presidents of the United States. Thus, to obtain the expected frequencies, simply multiply the sample size (n) by the relative frequency for each group (*p*):

Expected Number of Left-Handed Presidents: $f_e = (.10)(44) = 4.4$
Expected Number of Right-Handed Presidents: $f_e = (.90)(44) = 39.6$

Next, we determine the critical chi-square value. As with the oneway chi-square analysis, the degrees of freedom in the goodness of fit test is equal to k - 1, where 'k' is the number of groups in the sample. In our example, we have two groups in the sample: left-handed and right-handed Presidents. Thus, df = 2 – 1 = 1 and I will select an alpha level of α = .05. The formula for calculating the chi-square statistic is the same as that used in the oneway chi-square in Section 20.3; and as before, you subtract the expected frequency for group from the observed frequency for that group. You then square the difference and then divide the squared difference by the expected frequency for that group, before finally adding the quotients. This is done in the table below:

Group	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
Left-Handed Presidents	8	4.4	-3.6	12.96	2.945
Right-Handed Presidents	36	39.6	3.6	12.96	0.327
					$\chi^2 = 3.272$

From Table 5, the p-value associated with $\chi^2 = 3.272$ is *p* = .0736, which is greater than the selected alpha-level (.05); hence, we retain the null hypothesis and make no decision on the alternate hypothesis. Thus, the observed frequencies of left-handed and right-handed Presidents of the United States do not significantly deviate from the relative frequencies of left-handed and right-handed people in the population of the United States. Surprisingly, this is a good thing! It means that the groups within a population are not over or underrepresented in your sample.

20.5 Chi Square with Two Independent Variables

Chi-square analyses can also be used to examine whether a distribution of observed frequencies differs from a distribution of expected frequencies among the combinations of two independent variables. For example, in the previous section say I counted the number of males and the number of females in each of

the six college majors. I would have two nominal independent variables, College Major, and Sex, with a total of 6 (College Major) x 2 (Sex) = 12 mutually exclusive groups. I could count the number of students belonging to each of these 12 groups, and then determine the expected number of students who should belong to each of these groups, in order to perform an chi-square analysis.

When you have two nominal independent variables and the dependent variable is frequency, you perform a **factorial chi-square** to analyze the frequency data. This type of chi-square analysis will test for the presence of a statistically significant relationship between the two independent variables. For example, from the preceding paragraph, if a chi-square analysis performed on that set of observed frequencies was found to be statistically significant, it suggests that there is a relationship between a student's sex and their choice of college major that leads a student to enroll in Sensation & Perception.

When examining the relationship between two nominal independent variables and when the dependent variable is frequencies, you start by setting up a contingency table that lists all of the observed frequencies for each of the combinations created by the independent variables. Recall from Chapter 10 (Probability), a contingency table lists two independent variables, one represented by the rows and the other represented by the columns in the table. The table lists the observed frequency associated with each combination of the two independent variables. From such a contingency table, we will calculate the expected frequency for each combination of variables and then perform a chi-square analysis. Using the sex and major example from above, the contingency table might look the one below:

	Major						
Sex	**Psychology**	**Counseling & Human Services**	**Neuroscience**	**Occupational Therapy**	**Communications**	**Other**	**Row Totals**
Male	fo = 4	fo = 0	fo = 2	fo = 0	fo = 2	fo = 3	11
Female	fo = 13	fo = 3	fo = 1	fo = 6	fo = 2	fo = 3	28
Column Totals	17	3	3	6	4	6	$n_T = 39$

The row totals and the column totals are **marginal frequencies**. It is important to list these in the contingency table, because they are used to calculate the expected frequencies for each Sex by Major combination. A factorial chi-square can take one of two forms; however, both use exactly the same procedures and have the same assumptions. The only difference is how the frequency data is collected and how results are interpreted.

A **chi-square test for independence** is where the marginal frequencies vary independently, that is, the marginal frequencies for both independent variables are not fixed, and are unknown prior to collecting data. A chi-square test for independence is used to determine whether there is a relationship between the independent variables. The example above is a chi-square test for independence, because there is no way to know the marginal frequencies for either independent variable before the course had started.

A **chi-square test for homogeneity of variance** is where the marginal frequencies of <u>only one</u> independent variable are determined before data is collected (marginal frequencies for both independent variables cannot be fixed). For example, I may be interested in the choice of college major for male and female students that get their coffee at Starbucks. I could wait outside the Starbucks in the student center and ask the first 50 males and the first 50 females about their choice of major. In this example, I decided on the marginal frequencies of males and females before collecting data; that is, I decide to collect data from <u>only</u> 50 males and <u>only</u> 50 females. In this case, the marginal frequencies of males and females are fixed and cannot vary. A chi-square test for homogeneity of variance evaluates whether the difference between the observed frequencies and expected frequencies are equal (homogeneous) for males and females. For present purposes I cover only the chi-square test for independence. Just note that the testing procedures are exactly the same, the only difference is how the data is being collected and interpreted.

A factorial chi-square analysis has the same assumptions as the oneway chi-square: (1) The dependent variable is frequency; (2) the independent variables are on a nominal scale, and (3) the levels of the independent variables are mutually exclusive. A factorial chi-square has one additional assumption: the frequency of each group is at least five. Unfortunately, the frequencies of several groups in the sex and major example are less than five, thus, the example violates this important assumption.

A good example of where the chi-square test could be used to assess frequency data is congressional voting records from the US House of Representatives. Specifically, chi-square tests can examine the relationship between political party and voting ("Yea" and "Nay") on a particular bill. Such a situation has two independent variables, Political Party and Voting Decision, and the dependent is the frequency of responses in each combination of political party and voting decision.

One particular bill was House of Representatives Bill HR6, the *Energy Independence and Security Act* which passed in the House on December 18th, 2007. The bill set to "reduce our Nation's dependency on foreign oil by investing in clean, renewable, alternative energy resources, promoting new energy technologies, developing greater efficiency, and creating a Strategic Energy Efficiency and Renewable Reserve to invest in alternative energy, and for other purposes." Below, is a contingency table that lists the frequency for each combination of political party (Democrats and Republicans) and Decision on the bill (Yea, Nay, Abstain):

	Political Party		
Voting Decision	**Democrats**	**Republicans**	**Row Totals (RMF)**
Yea	$f_0 = 219$	$f_0 = 95$	314
Nay	$f_0 = 4$	$f_0 = 96$	100
Abstain	$f_0 = 9$	$f_0 = 10$	19
Column Totals (CMF)	232	201	$n_T = 433$

Note that n_T should be equal to the sum of the row marginal frequencies (314 + 100 + 19 = 433) and also be equal to the sum of the column marginal frequencies (232 + 201 = 433). I admit that the observed frequency of Democrats voting "Nay" is less than five; hence, the assumption that each cell frequency is at least five is violated, but we'll ignore that for now, because 4 is close to 5.

The first step is to determine the critical chi-square value. Degrees of freedom in a factorial chi-square are equal to df = (c – 1)(r – 1), where 'c' is the number of columns in the contingency table (number of levels for the column independent variable) and 'r' is the number of rows in the contingency table (number of levels for the row independent variable). Here, Political Party (column variable) has c = 2 two levels and Voting Decision (row variable) has r = 3 levels. Thus, there are (2 – 1)(3 – 1) = (1)(2) = 2 degrees of freedom and I will use an alpha-level of α = .05.

Next, calculate the expected frequency for each combination of the independent variables. In the case of the factorial chi-square analysis, a single f_e that will be equal for all cells will not suffice; rather, a different f_e value is necessary for each cell. Each f_e value will be based on the **row marginal frequency** (**RMF**) and **column marginal frequency** (**CMF**) associated with a particular cell (group). The expected frequency of any cell can be calculated using the following equation:

$$f_e = \frac{(CMF)(RMF)}{n}$$

RMF and CMF refer to the row marginal frequency and column marginal frequency associated with a particular group, respectively. For example, the observed frequency of Democrats voting Yea is $f_o = 219$. The row marginal frequency that is associated with that cell in the table is RMF = 314, and the column marginal frequency associated with that cell is CMF = 232. That is, to figure out the expected frequency of Democrats voting Yea, you locate that cell in the contingency table, and then use the CMF of that cell's column and the RMF of that cell's row in the formula above. Using the equation above, the expected frequency for Democrats voting Yea is:

$$f_e = \frac{(232)(314)}{433} = 168.24$$

If the independent variables Political Party and Voting Decision were unrelated, we would expect about 168 Democrats to vote Yea on this bill. In the table below, the expected frequencies are listed with the observed frequency for each combination of the two independent variables. Although the marginal frequencies are not listed, note that within rounding error the expected frequencies in a row sum to the row marginal frequency and the expected frequencies in a column sum to the column marginal frequency:

	Political Party	
Voting Decision	**Democrats**	**Republicans**
Yea	$f_o = 219$ $f_e = 168.240$	$f_o = 95$ $f_e = 145.760$
Nay	$f_o = 4$ $f_e = 53.580$	$f_o = 96$ $f_e = 46.420$
Abstain	$f_o = 9$ $f_e = 10.180$	$f_o = 10$ $f_e = 8.820$

Once the expected frequencies have been obtained the next step is to perform the chi-square test: The formula for doing so and the procedure is identical to the chi-square test for a one-way design. With a factorial chi square, each expected frequency is compared to each observed frequency. The chi-square procedure is demonstrated in the table below:

Group	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
Democrats/Yea	219	168.240	50.760	2576.578	15.315
Democrats/Nay	4	53.580	-49.580	2458.176	45.879
Democrats/Abstain	9	10.180	-1.180	1.392	0.137
Republican/Yea	95	145.760	-50.760	2576.578	17.677
Republican/Nay	96	46.420	49.580	2458.176	52.955
Republican/Abstain	10	8.820	1.180	1.392	0.158
					$\chi^2 = 132.121$

From Table 5, the p-value associated with $\chi^2 = 132.121$ (use 40.00 in Table 5) is $p < .0001$, which is less than the selected alpha-level (.05); hence the null hypothesis is rejected and the alternate hypothesis is accepted. Thus, we would conclude that for this particular House Bill there is a significant relationship between a congresspersons political party and voting decision. As before, we can describe the relationship between the independent variables by examining first the $(f_o - f_e)^2/f_e$ values and then the $f_o - f_e$ values, for each group. The two largest $(f_o - f_e)^2/f_e$ values are Democrats Voting Nay (45.879) and Republicans Voting Nay (52.955). Examining the $f_o - f_e$ values, Democrats were much *less* likely to vote Nay on this bill (-49.580); whereas Republicans were much *more* likely to vote Nay on the Bill (49.850). The next two largest $(f_o - f_e)^2/f_e$ values are Democrats Voting Yea (15.315) and Republicans Voting Yea (17.677). Examining the $f_o - f_e$ values, Democrats were much *more* likely to vote Yea on this bill (50.760); whereas Republicans were much *less* likely to vote Yea on the Bill (-50.760).

20.6 Strength and Power of Chi Square

The strength of the relationship between the two nominal variables that were analyzed with a chi-square test can be estimated by **Cramer's Contingency Coefficient (C)**. This value is similar to eta-squared (η^2), the proportion of variance explained in a relationship between variables. Cramer's C is used only when you have a two-way chi-square and the number of levels of at least one variable is greater than two. Cramer's C is calculated from:

$$C = \sqrt{\frac{\chi^2}{n+\chi^2}} = \sqrt{\frac{\square^2}{\square+\square^2}}$$

The chi-square values in the expression are the obtained chi-square statistics. The strength of the relationship for the example in the preceding section is:

$$C = \sqrt{\frac{132.121}{433+132.121}} = 0.484$$

Thus, the amount of variance that is explained in the relationship between political party and voting decision on this Bill is about 48.4%. A more standardized measure of effect size is Cohen's *w*. which can be calculated from Cramer's C:

$$w = \sqrt{\frac{C^2}{1-C^2}} = \sqrt{\frac{0.484^2}{1-0.484^2}} = 0.552$$

Cohen provides descriptive labels for this effect size:

Effect Size	Cohen's w
"Small"	.10
"Medium"	.30
"Large"	.50

In this case, we have a "large" effect size. Cramer's w is also useful for determining the statistical power that was observed in a chi-square test. Using G*Power 3, under Test Family select "χ^2 tests," and under Type of Power Analysis select "Post hoc:...". Enter the information from the chi-square test into the appropriate areas under Input Parameters (*w* = .0552, α = .05, Total sample size = 433, df = 2). G*Power tells us that Power is equal to 1.000, which simply means that the probability that we correctly rejected a false null hypothesis is greater than 0.999.

20.7 Special Case of the 2x2 Chi-Square

The chi-square test is not terribly difficult to perform; however, it can be time consuming. There is a special case where a *computational formula* for the chi-square test can be used. When you have a factorial design and there are only two independent variables and each independent variable has only two-levels, you can use the following computational formula, which is described below:

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Note, that this can be used only if you have a 2 x 2 design; if you have anything larger than a 2x2 design, you must use the procedures discussed earlier. The letters, a – d, in the formula refer to specific cells within a 2 x 2 contingency table that is formed between the two independent variables. The cells associated with each letter are listed in the table below. The 'n' is total frequency:

	Independent Variable "A"	
Independent Variable "B"	A_1	A_2
B_1	a	b

B_2	c	d

As an example, say that we count the number of male republicans, male democrats, female republicans, and female democrats in a political science course and obtain the frequencies in the table below:

	Political Party	
Sex	Democratic	Republican
Males	f_o = 30	f_o = 30
Females	f_o = 20	f_o = 10

In this example there are n = 100 students taking the course and there are two independent variables (Sex and Political Party), each with two levels. This is a 2 x 2 factorial design. You could use the procedures that were developed in section 20.5 to perform a chi-square analysis on these data, or, you could use the computational formula. Plugging in the values associated with each cell in the example above into the 2 x 2 chi-square expression, we have:

$$\chi^2 = \frac{100(30\times10-30\times20)^2}{(30+30)(20+10)(30+20)(30+10)}$$

Once you have all of the values plugged into the equation the rest is basically just appropriate application of the order of operations:

$$\chi^2 = \frac{100(30\times10-30\times20)^2}{(30+30)(20+10)(30+20)(30+10)} = 2.5$$

Because there are two rows and two columns in the 2 x 2 design there is only 1 degree of freedom. From Table 5, the p-value associated with $\chi^2 = 2.5$ is $p = .1138$, which is greater than the selected alpha-level (.05); hence the null hypothesis is retained and we conclude there is not enough evidence to suggest a relationship between sex and political party among students taking a political science course.

Effect size (strength of the relationship) can also be measured for a 2 x 2 chi square, but instead of using Cramer's C when the design is 2 x 2 the **phi-coefficient** (**Φ**) should be used. The phi-coefficient is calculated by taking the square root of the quotient of the obtained chi-square over the total frequency:

$$\Phi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{2.5}{100}} = 0.158$$

CH 20 Homework Questions

1. When is a chi-square test typically used to analyze a bivariate relationship?

2. What is the difference between the chi-square test of independence and the chi-square test of homogeneity?

3. Describe the difference between the observed frequencies and the expected frequencies.

4. State the critical value of χ^2 for a chi-square test for an alpha level of $\alpha = .05$ and also of $\alpha = .01$ under each of the following conditions:

a. $k = 2$
b. $k = 3$
c. $k = 4$
d. $r = 2, c = 3$
e. $r = 3, c = 3$
f. $r = 2, c = 4$
g. $r = 4, c = 4$

5. Determine the p-value for each of the following given scenarios:

a. $k = 2, \chi^2 = 4.50$
b. $k = 2, \chi^2 = 2.60$
c. $k = 6, \chi^2 = 11.00$
d. $k = 6, \chi^2 = 9.00$
e. $r = 2, c = 2, \chi^2 = 3.50$
f. $r = 3, c = 3, \chi^2 = 12.20$
g. $r = 2, c = 4, \chi^2 = 6.40$

6. *Use the following to answer the questions below*: A psychology professor is interested in the numbers of students in each college class (freshmen, sophomore, junior, senior) enrolled in his Fundamentals of Psychology lecture and wants to know whether there is significant difference across the numbers of students in each a college class. The professor records the class of each student who is enrolled in his Fundamentals of Psychology course. The data follows.

College Class			
Freshmen	Sophomore	Junior	Senior
$f_o = 162$	$f_o = 39$	$f_o = 6$	$f_o = 5$

a. Calculate the expected frequency of each college class.
b. Perform a chi-square test on the data.

c. What is the p-value for this test statistic?
d. Assuming α = .05, is there a statistically significant difference across the four college classes, that is, is there is difference between the observed and expected frequencies? What decisions are made with the hypotheses?
e. Use the $(f_o - f_e)^2/f_e$ and $(f_o - f_e)$ values to discern the significant result.

7. *Use the following to answer the questions below*: You are interested in the mating habits of baboons. You find a baboon troop with eight males of breeding age and eighty females of breeding age. You assume each male has equal mating rights to the females; that is, each male baboon will mate with an equal number of female baboons. You observe the baboon troop during mating season find that each male baboon mates with the following number of female baboons:

Male 1	Male 2	Male 3	Male 4	Male 5	Male 6	Male 7	Male 8
1	5	0	1	3	65	2	3

a. Based on this information, how many female baboons is each male expected to mate with?
b. Perform a chi-square test on the data.
c. What is the p-value for this test statistic?
d. Assuming α = .05, is there a statistically significant difference in the number of female baboons that mate with each male baboon? What decisions are made with the hypotheses?

8. *Use the following to answer the questions below*: The data from a recent US Congressional vote are displayed below by the political party of the congressional representative and his/her vote. Use this data to answer the questions that follow.

	Decision			
Political Party	Yea	Nay	Abstain	Totals
Democratic	f_o = 140	f_o = 30	f_o = 10	180
Republican	f_o = 20	f_o = 225	f_o = 10	255
Totals	160	255	20	n_T = 435

a. Calculate the expected frequency of each cell.
b. Perform the chi-square test on these data.
c. What is the p-value for this test statistic?
d. Assuming α = .05, is there a statistically significant difference between the observed and the expected frequencies? What decisions are made with the hypotheses?
e. Use the $(f_o - f_e)^2/f_e$ and $(f_o - f_e)$ values to discern the significant result.
f. Calculate Cramer's Contingency Coefficient.
g. Calculate Cohen's w from Cramer's Contingency Coefficient
h. Using G*Power 3, how much statistical power was there to reject the null hypothesis?

9. *Use the following to answer the questions below*: The table displays the observed frequencies of political party affiliation and residential area for a sample of individuals.

	Residential Area			
Party Affiliation	City	Suburbs	Country	Totals

Democrat	120	80	60	260
Republican	40	80	20	140
Independent	40	40	20	100
Totals	200	200	100	$n_T = 500$

a. Calculate the expected frequency of each cell.
b. Perform the chi-square test on these data
c. What is the p-value for this test statistic?
d. Assuming α = .05, is there a statistically significant difference between the observed and the expected frequencies? What decisions are made with the hypotheses?
e. Use the $(f_o - f_e)^2/f_e$ and $(f_o - f_e)$ values to discern the significant result.
f. Calculate Cramer's C statistic.
g. Calculate Cohen's w from Cramer's Contingency Coefficient.
h. Using G*Power 3, how much statistical power was there to reject the null hypothesis?

10. *Use the following to answer the questions below*: There is a stereotype that university professors tend to be atheist or agnostic. This stereotype is even more pronounced if the professor has an appointment in the sciences (e.g., biology, chemistry, physics, psychology) compared to the humanities (e.g., history, philosophy, theology, English). To examine whether there is a relationship between religions view and appointment of a professor, 250 professors are asked whether their appointment is in the sciences, the humanities, or some other field and whether they are Religious, Agnostic, or Atheist. The following frequencies are observed:

Field	Religious	Agnostic	Atheist
Sciences	25	25	75
Humanities	25	25	50
Other	15	5	5

a. Calculate the expected frequency of each cell.
b. Perform the chi-square test on these data
c. What is the p-value for this test statistic?
d. Assuming α = .05, is there a statistically significant difference between the observed and the expected frequencies? What decisions are made with the hypotheses?
e. Calculate Cramer's C statistic.
f. Calculate Cohen's w from Cramer's Contingency Coefficient.
g. Using G*Power 3, how much statistical power was there to reject the null hypothesis?

11. *Use the following to answer the questions below:* Suppose the relative frequencies of students belonging to each of three colleges at a particular university are as follows:

College	rf
Natural Sciences	.20
Business School	.25
Social Sciences	.35

Arts and Humanities	.20

Furthermore, suppose that a general education course of 100 students yields the following observed frequencies:

College	f_o
Natural Sciences	40
Business School	20
Social Sciences	20
Arts and Humanities	20

a. Determine the expected frequencies for each college.
b. Perform a chi-square goodness of fit test on these data.
c. What is the p-value for this test statistic?
d. Assuming α = .05, is there a statistically significant difference between the observed and the expected frequencies? What decisions are made with the hypotheses?
e. Use the $(f_o - f_e)^2/f_e$ and $(f_o - f_e)$ values to discern the significant result.

12. *Use the following to answer the questions below:* Suppose that the relative frequencies of male and female students at a particular high school are as follows:

College	rf
Male	.47
Female	.53

Suppose that a course of 50 students yields the following observed frequencies:

College	f_o
Male	26
Female	24

a. Determine the expected frequencies for each college.
b. Perform a chi-square goodness of fit test on these data.
c. What is the p-value for this test statistic?
d. Assuming α = .05, is there a statistically significant difference between the observed and the expected frequencies? What decisions are made with the hypotheses?

13. *Use the following to answer the questions below:* A political science teacher is examining the relationship between political party affiliation and college class among the students in one of his lectures. The following contingency table was created:

	Political Party Affiliation	
Sex	Democratic	Republican

Underclassmen	10	20
Upperclassmen	25	5

a. Using the computational formula for a 2x2 contingency table, calculate the obtained value of the chi-square statistic.
b. What is the p-value for this test statistic?
c. Assuming α = .05, is there a statistically significant difference between the observed and the expected frequencies? What decisions are made with the hypotheses?
d. Calculate the phi coefficient.

14. I want to know whether there is a relationship between sex (males/female) and getting caught text-messaging in class. For one academic year I count the number of males and females caught text-messaging in my classes. I also count the numbers of males and females from those classes that I did not catch text-messaging (i.e., who were not likely text-messaging). The frequencies of males and females caught text-messaging and not caught text-messaging in class for this academic year are below. Use this information to answer the following questions:

	Caught Text Messaging	
Sex	Yes	No
Males	15	55
Females	30	80

a. Using the computational formula for a 2x2 contingency table, calculate the obtained value of the chi-square statistic.
b. What is the p-value for this test statistic?
c. Assuming α = .05, is there a statistically significant difference between the observed and the expected frequencies? What decisions are made with the hypotheses?
d. Calculate the phi coefficient.

15. For each of the following situations, use G*Power 3 to find the <u>total</u> number of subjects that would be needed to achieve the desired level of Power.

a. w = .10, α = .05, Power = .80, df = 2
b. w = .20, α = .05, Power = .80, df = 2
c. w = .10, α = .05, Power = .95, df = 2
d. w = .20, α = .05, Power = .80, df = 4

16. For each of the following situations, use G*Power 3 to find the amount of Power, based on the parameters given.

a. w = .30, α = .01, n = 100, df = 2
b. w = .30, α = .01, n = 100, df = 4
c. w = .40, α = .01, n = 200, df = 2
d. w = .40, α = .05, n = 50, df = 2

Chapter 21: Power Analysis and Replication

21.1 Review of Effect sizes and Power

Earlier chapters addressed issues related to effect size and power as well as factors that influence power, such as sample size, alpha level, directionality, and type of inferential test. To recap, effect size is the strength of the relationship being examined and statistical power is the probability of correctly rejecting a null hypothesis (i.e., of correctly detecting a statistically significant result). Effect size and power are related, in that, to detect small effect sizes more statistical power is required, whereas to detect larger effect sizes less power is needed. Sample size is also related to power, where larger sample sizes generally increase the power to detect an effect.

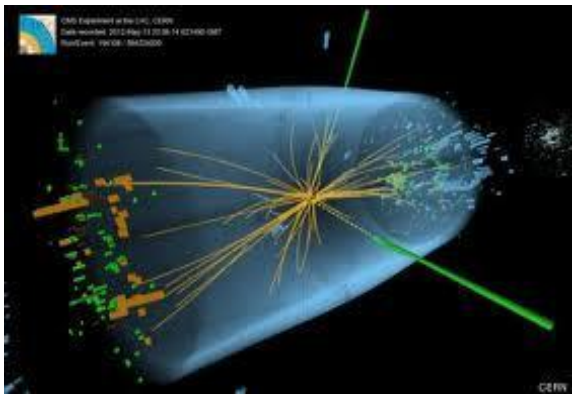
Recall, power can be used in one of two ways: (1) After an inferential test is performed on a set of data, power can be calculated based on observed results. This **post-hoc power analysis** is used to estimate the observed power of an inferential test and to ensure the probability of correctly rejecting a null hypothesis meets some accepted criterion, which is generally .80 or greater. (2) Before beginning data collection a power analysis is performed to determine needed sample size. This is an **a priori power analysis** and is used during the planning phase of a study to determine the number of subjects needed to achieve a certain level of statistical power based on a predicted effect size.

A priori power analyses are also important when planning **replication studies**, which are studies that either fully or partially replicate the methods of a previously published study. **Full replications** use the exact same methods, stimuli, and procedures from a previous study, thus, full replications are studies in which a study is performed again. **Partial replications** mimic certain aspects of a previous study, but change some elements, with the intent to extend the results beyond the previous study. The reasons for partial replication are easy to identify, that is, to extend previous results. What is the purpose of a full replication? Why would researchers conduct someone else's study? This is a valid question, because replication studies are less likely to be published, so why do people do them?

The simple answer is replication can provide confirmation that an observed result is real and not a fluke. Remember, when you claim a result is statistically significant, say $p = .01$ and a statistical power of .80, you are saying there is a 1% chance of detecting your result even if the null hypothesis was true and an 80% chance you have correctly rejected a false null hypothesis. Thus, there is a chance your result was obtained with the null being true (i.e., a Type I error). Remember, inferential statistics are based on probabilities, so there's always a chance of being wrong. If a study's results are replicated, this makes it more likely a result was real and was not just a fluke. Two examples will help make the point.

Social psychologist Daryl Bem published a controversial paper in the *Journal of Personality and Social Psychology*, in which he claimed to show evidence of the *psi phenomenon* (*extrasensory perception* or *ESP*; Bem, 2011). Such a claim is great, as many people doubt the existence of the psi phenomenon. To his credit, Bem conducted a priori power analyses based on an expected effect size and the conventional level of statistical power (.80), so he did collect samples of sufficient size and in nine separate experiments consistently demonstrated evidence of the psi phenomenon. Because of the controversy surrounding the psi phenomenon researchers attempted to replicate Bem's study, but, after several published replications (e.g., Ritchie, Wiseman & French, 2012) and mentions of replication studies by other researchers, the evidence of psi declined. Specifically, each replication attempt failed to replicate

Bem's results, suggesting that the original results may be a fluke (Type I error).[1]



As a second example, The European Organization of Nuclear Research (CERN) recently announced the discovery of the Higgs hoson; a subatomic particle that interacts with other subatomic particles to create matter to create mass. The Higgs boson was predicted in the standard model of physics in the 1960's and, until CERN's announcement, was only theoretical. Although this is a monumental discovery, before concluding the data collected by CERN is evidence of the Higgs boson additional data must be collected; thus, the study and the results must be replicated, because it is possible, though highly unlikely, the original evidence for the Higgs boson was a fluke. The evidence must be replicated again and again to show that the boson was actually discovered.

In short, replication is important to science for verifying or disconfirming new discoveries. Importantly, when preparing for replications, statistical power analyses must be conducted to ensure that the original study is repeated in as many ways as possible, which includes using appropriate sample sizes.

21.2 Preparing for Replication

There is more to conducting a replication study than making sure the same procedures, equipment, materials, and population are used; hence, replication is more than just methodology. To properly conduct a replication study you must also ensure the study will have sufficient statistical power to detect an effect of the size reported in the original study. A power analysis must be performed in order to determine the appropriate number of subjects to test in the replication study so the eventual statistical decision will be justified based on the data that is collected. Thus, a priori power analyses are all about determining the appropriate sample size to use. Here are the general steps involved with replicating a study:

1. Identify a study to be replicated.
2. Develop the methods of the replication study, which come from the methods of the original study..
3. Perform a power analysis to determine the appropriate sample size. More on this below.
4. Begin the replication study.

There are many power analysis tools available on the Internet, but G*Power 3, which was introduced in the earlier chapters, is quite useful (and free!). To perform a priori power analyses you need to know several values from the study you are planning to replicate. Specifically, to properly replicate a study you must know:

1. The type of inferential test to be used.
2. The alpha level used (if unknown or if the p-values were reported, it's OK to use .05).
3. Whether the inferential test was one-tailed or two-tailed (if applicable).
4. The effect size in the original study (e.g., Cohen's *d*, Cohen's *f*, η^2).

Most of these can be obtained easily from the Results sections of a published paper, but if not contacting the author may be needed. One issue is many studies do not report effect size measures with the results of an inferential test, and this is especially true in older studies. For example, a paper may report the result

[1] For interesting commentary on this issue, see Chris French's blog at The Guardian: http://www.guardian.co.uk/science/2012/mar/15/precognition-studies-curse-failed-replications HYPERLINK "http://www.guardian.co.uk/science/2012/mar/15/precognition-studies-curse-failed-replications"

of an ANOVA as, $F(3, 36) = 5.00$, $MSE = 10$, $p < .05$. This provides information about the statistical significance of the ANOVA, but does not provide any direct information about the effect size, because Cohen's *f* and η^2 were not reported. If the results of the ANOVA included Cohen's *f* or η^2, those values could be used in G*Power to determine the appropriate sample size for a replication study. But in cases where effect size measures are not reported with an inferential test, the effect size must be estimated from the available information. The next section describes some methods for doing so.

21.3 Effect Size Computations

As mentioned in the preceding section, one problem in conducting a priori power analyses for replication studies is determining the effect size in a published study, because authors often fail to report effect size measurements (Cohen's *d*, Cohen's *f*, Cohen's *w*, η^2) with the results of inferential statistical tests. Thus, the *F* statistic for ANOVA is reported, but a corresponding value for η^2 or Cohen's *f* is not. Luckily, there are several straightforward methods for calculating the effect size from the information that is almost always reported in any inferential statistic. This section covers some ways for computing effect size measurements from available information in Results sections. The one thing to keep in mind is that when you need to calculate an effect size for the results of a published study, identify all of the information that is available in that study first and then choose the best method for calculating the appropriate effect size.

Independent Groups t-tests

Assume a researcher compared two independent groups (A_1 and A_2) with $n = 10$ subjects per group using an independent groups *t*-test and obtained the following results: $M_{A1} = 10$, $SD_{A1} = 3$; $M_{A2} = 14$, $SD_{A2} = 4$; $t(18) = 2.53$, $SE = 1.58$, $p < .05$ (two-tails). The standard measure of effect size, which is also used in G*Power, is Cohen's *d*, which is defined as the number of standard deviations that separate two means and is calculated thus:

$$d = \left| \frac{\overline{X}_1 - \overline{X}_2}{\widehat{s}_{Pooled}} \right|$$

Although it is almost certainly the case that the mean of each condition, or at least the mean difference between the two conditions, will be reported in a published article, it is almost never the case that the pooled standard deviation ($\widehat{\sigma}_{pooled}$) is reported. Thus, the pooled standard deviation may need to be estimated from available information. Recall that the pooled standard deviation is simply the square root of the pooled variance: $\widehat{\sigma}_{pooled} = \sqrt{\widehat{\sigma}^2_{pooled}}$, and the pooled variance can be found from this formula:

$$\widehat{s}^2_{pooled} = \frac{(n_1 - 1)\widehat{s}^2_1 + (n_2 - 1)\widehat{s}^2_2}{n_1 + n_2 - 2}$$

Thus, the pooled standard deviation is:

$$\widehat{s}^2_{pooled} = \sqrt{\frac{(n_1 - 1)\widehat{s}^2_1 + (n_2 - 1)\widehat{s}^2_2}{n_1 + n_2 - 2}}$$

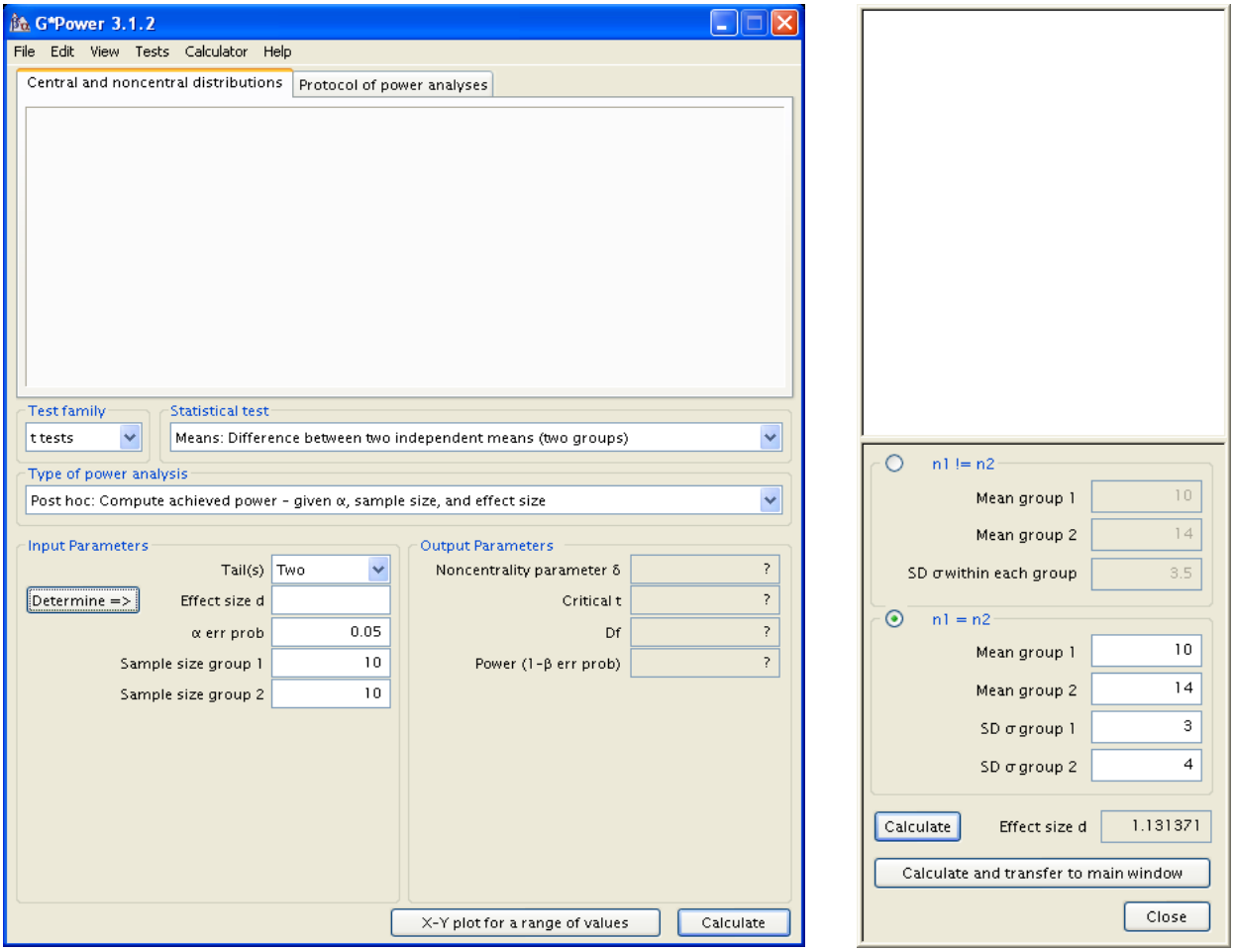
From the information in the example above, we know that n_1 and n_2 are both 10, that is, the sample size for each condition was 10. The $\widehat{\sigma}^2_1$ and $\widehat{\sigma}^2_2$ values are the variances for each condition. Although we do not know the variance for each condition, we do know the standard deviation for each condition ($SD_{A1} = 3$ and $SD_{A2} = 4$), and because we know that the variance is the square of the standard deviation, we can easily determine the variance of the two conditions: $\widehat{\sigma}^2_1 = 3^2 = 9$ and $\widehat{\sigma}^2_2 = 4^2 = 16$. If we take the known sample sizes and variances from each group we can plug those values into the formula above to solve for the pooled standard deviation:

$$\widehat{s}^{\,2}_{pooled} = \sqrt{\frac{(10-1)9+(10-1)16}{10+10-2}} = 3.54$$

Using this value and the means of the two conditions (M_{A1} = 10 and M_{A2} = 14), Cohen's *d* is equal to:

$$d = \left|\frac{10-14}{3.54}\right| = 1.13$$

This value could then be used in a power analysis to determine the appropriate sample size to use in a replication study. Incidentally, G*Power has a built in feature for determining the effect size based on the means and standard deviations of conditions. With G*Power open and either "Post hoc" or "A priori" selected under Type of power analysis, click the Determine => button, and a window will open to the right:



In the window that opens you can enter the mean and standard deviation of each condition (under n1 = n2) and click the Calculate button to obtain the effect size, which you can then transfer to the main window by clicking the 'Calculate and transfer to main window' button.

It may be the case the standard deviations are not reported, but the standard error of the mean is reported. Recall the standard error of the mean is equal to the standard deviation divided by the square root of the sample size. In the example above the standard errors are SE_{A1} = 3/√10 = 0.949 and SE_{A2} = 4/√10 = 1.265. If instead of the standard deviation the standard error of the mean was reported, you can calculate the standard deviation from the following formula:

$$\widehat{s} = SE\sqrt{n}$$

Once the standard deviations have been calculated from the standard errors, you can use enter that information into G*Power to determine the effect size. If the standard deviations and standard errors are not reported, the above procedures will not work, because the standard error, the variance, or the standard deviation is needed to determine the effect size. In those situations, there are computational formulas that can be used to estimate the effect size from a *t*-test directly. In cases where the *t*-statistic and the sample size of each condition is known, the following formula can be used to estimate Cohen's *d*:

$$d = \frac{2t}{df}$$

Here, *t* is the value of the independent-groups *t*-test, which was equal to 2.53 in this example. Importantly, this formula cannot be used for correlated-samples *t*-tests, as these formulas overestimate the effect size for paired samples tests. I provide some tops for calculating power of paired samples *t*-tests later. Plugging in the *t*-value and the degrees of freedom into this formula, we get:

$$d = \frac{2x2.53}{18} = 1.192$$

Although slightly larger than the *d* = 1.13 values obtained above and through G*Power, it is quite close.

Paired Samples *t*-tests

There is some controversy about how to calculate effect size in a paired samples t-test. The following formula cannot be used, because it overestimates the true effect size:

$$d = \frac{2t}{\sqrt{df}}$$

Recall from Chapter 14, the following was used to calculate Cohen's *d* for the paired-samples t-test:

$$d = \frac{M_D}{s_{avg}}$$

Where s_{avg} is the average standard deviation between the two paired samples:

$$s_{avg} = \sqrt{\frac{\hat{s}_1^2 + \hat{s}_2^2}{2}}$$

Calculating Cohen's d in this way is easily done if you have access to the raw data. Or, if the standard deviations are reported in a paper the variance can be calculated by squaring the standard deviations, you can then use the variance to calculate s_{avg}. Unfortunately, when the variances or standard deviations are not reported, you must rely on other means to calculate Cohen's d.

Recall though that Cohen's *d* (d_z in the paired samples *t*-test) is the difference between two means divided by the pooled standard deviation. The pooled standard deviation can also be through of as the **standard deviation of the difference** ($\hat{s}_D$). In the paired samples *t*-test, the test statistic is, technically speaking, equal to the mean difference (M_d) divided by the standard error of the difference (SE_d):

$$t = \frac{M_D}{\hat{s}_{\overline{D}}}$$

The standard error of the difference is equal to the standard deviation of the difference divided by the square root of the sample size:

$$\hat{s}_{\overline{D}} = \frac{\hat{s}_D}{\sqrt{n}}$$

Importantly, the standard error of the difference is the term that is almost always reported with the results of a correlated-samples *t*-test. Also, if you know the standard error of the difference and the sample size, you can calculate the standard deviation of the difference from the following formula:

$$\hat{s}_D = \hat{s}_{\overline{D}}\sqrt{n}$$

For example, say a researcher compared two means (B_1 and B_2) in a paired samples *t*-test and obtained the following results: $M_{B1} = 6$; $M_{B2} = 9$; $t(9) = 2.50$, $SE = 1.20$, $p < .05$ (two-tailed). The *SE* value (1.20) is the standard error of the difference (SE_d in the formulas above), which can be used to calculate the standard deviation of the difference:

$$\hat{s}_D = 1.20\sqrt{10} = 3.794$$

Once you know the standard deviation of the difference, you can use that to calculate Cohen's *d*, by using $\hat{s}_p$ for the pooled standard deviation:

$$d = \frac{6-9}{3.794} = 0.791$$

Analysis of Variance

The following is mostly applicable to finding effect size in oneway between subjects ANOVA, but the techniques could also be used to calculate effect size for other ANOVA techniques. As with *t*-tests, it is often the case that the results of an *F*-test for ANOVA include the *F*-statistic, the degrees of freedom, the *p*-value, and the mean square for the error term, but not a measure of effect size (Cohen's *f*, η^2) and, hence, must be estimated from available information. Recall that the proportion of explained variance, η^2 is found from the formula:

$$\eta^2 = \frac{SS_{Effect}}{SS_{Error}}$$

And the effect size, Cohen's *f* is found from the formula:

$$f = \sqrt{\frac{\eta^2}{1-\eta^2}}$$

Thus, except for alternative examples of Cohen's *f* below, to obtain *f* you need to first calculate η^2. But, the sums of squares are almost never reported in published articles; hence, the formula above is useful only when one has first-hand knowledge of the data. But η^2 can also be thought of, more correctly, as the following:

$$\eta^2 = \frac{SS_{Effect}}{SS_{Efect} + SS_{Within}}$$

Where SS_{Effect} is the sum of squares for the effect you are interested in and SS_{Effect} is the sum of squares for the error term. But again, you need to know the sums of squares values, which is rarely the case. However, with some mathematical rearrangement and manipulation, this formula can be rewritten as:

$$\eta^2 = \frac{F \times df_{Effect}}{F \times SS_{Effect} + df_{Within}}$$

where df_{Effect} is the degrees of freedom between-groups (i.e., for the effect of the independent variable), df_{Effect} is the degrees of freedom within subjects (i.e., for the error term), and *f* is the *F*-statistic in ANOVA.

Source of Variance	SS	df	MS	F
Between	150	3	50	5
Within	360	36	10	
Total	510	39		

For example, say that a researcher runs a oneway between subjects ANOVA on a set of data with $k = 4$ levels of an independent variable and $n = 10$ subjects per level. Here is a hypothetical ANOVA summary table of the results. From the original formula for η^2, the proportion of explained variance is:

$$\eta^2 = \frac{150}{510} = 0.294$$

Using the modified formula from above, η^2 is:

$$\eta^2 = \frac{5(3)}{5(3)+36} = 0.294$$

Whichever formula is used, the resulting value can be used in the formula for Cohen's *f* and then used in G*Power. In this example, Cohen's *f* is:

$$f = \sqrt{\frac{0.294}{1-0.294}} = 0.645$$

Chi-Square Analyses

Although, like ANOVA and *t*-tests, it is rare for authors to report effect size measures with a chi-square test Cohen's *w*, Cramer's Correlation Coefficient [C], Φ), calculating these values from given information is relatively easy. Recall, that Cohen's *w* is computed using the following formula:

$$w = \sqrt{\frac{C}{1-C}}$$

Where C is Cramer's Correlation Coefficient, which is computed from the following formula:

$$C = \sqrt{\frac{\chi^2}{n+\chi^2}}$$

To calculate C you need the chi-square statistic and total number of subjects *n*), which are almost always reported in the results of an analysis using chi-square. For example, say an author reports the following results of a chi-square test of independence: $\chi^2(2, N = 1000) = 3.55$, $p < .05$. Cramer's C is equal to:

$$C = \sqrt{\frac{355}{1000+355}} = 0.059$$

From Cramer's C, you can now calculate Cohen's *w*:

$$w = \sqrt{\frac{0.059^2}{1-0.059^2}} = 0.055$$

This Cohen's *w* value would then be entered into G*Power to be used in a power analysis.

21.4. Examples

This section provides examples for conducting power analyses based from published studies. Note the procedures of power analyses are the same and the only thing that differs is how effect size is estimated.

Bem (2011) reported an experiment entitled "Retroactive Facilitation of Recall II" (Experiment 9), in which he observed greater "recall" of words prior to the words being studied. That is, subjects wrote down a greater percentage of words that they would see in the future compared to words they would not see in the future. The mean performance score of 4.21% was compared to 0% using a one-sample *t*-test, $t(49) = 2.96$, $p = .002$ (one-tail), $d = 0.42$. Thus, in this case, the effect size was provided. We want to replicate Bem's (2011) Experiment 9 and we know Cohen's $d = .42$ with a directional one-sample t-test. We choose an alpha level of .01 and want Power = .90. Using G*Power, we enter these and find that we need a sample size of n = 77 to achieve this power = .90 based on Cohen's *d* of .42 and an alpha level of .01.

As a second example, say that a researcher examined the difference in SAT math scores between males and females and found that males had significantly greater SAT math scores ($M = 670$) than females ($M = 640$), $t(98) = 2.10$, $SE = 500$, $p < .05$ (two-tails). In this case, the researcher had to use an independent groups *t*-test, because the two groups are naturally independent (assume the group sizes were equal, that is, $n = 50$ for both male and females). Because we do not know the standard deviations or the standard errors, we must estimate Cohen's *d* using the following formula:

$$d = \frac{2t}{\sqrt{df}}$$

From the information given, we know that in order to replicate this study we need to use a non-directional, independent groups *t*-test. We decide to use an alpha level of .01 and we want Power = .90. Using G*Power, we find that we need n = 342 subjects (171 per group).

As a final example, assume that a researcher conducts a study comparing three levels of orange juice consumption to examine orange juice consumption on number of colds over a five-year period. The researcher recruits *n* = 300 subjects, with 100 subjects per level of the independent variable. The data are analyzed using a oneway between subjects ANOVA and the results are, $F(2, 297) = 10.50$, $MSE = 5000.00$, $p < .05$. Because no effect size measures were provided by the author, Cohen's *f* must be estimated from the result of the *F*-test:

$$F = \sqrt{F \times \frac{df_{Effect}}{df_{Effect}}} = \sqrt{10.50 \times \frac{2}{297}} = 0.266$$

From the information given, we know that in order to replicate this study we need to use a oneway between subjects ANOVA with three groups. We decide to use an alpha level of .05 and we want Power = .80. Using G*Power, we find that we need n = 141 subjects (47 per group).

CH 21 Homework Questions

For #1 – 10, use the available information to calculate the effect size.

1. Independent groups *t*-test: $t(28) = 2.00$, $SE = 10.00$, $p < .01$ (one-tailed)
2. Paired samples *t*-test: $t(9) = 5.00$, $SE = 4.50$, $p < .05$ (one-tailed)
3. Independent groups *t*-test: $M_1 = 15$, $SD_1 = 5$, $n_1 = 20$; $M_2 = 18$, $SD_2 = 9$, $n_2 = 20$
4. Pearson correlation: $r(98) = .66$, $p < .01$ (two-tailed)
5. Oneway between subjects ANOVA: $F(1, 29) = 3.50$, $MSE = 100.00$, $p < .05$
6. Oneway between subjects ANOVA: $F(4, 95) = 4.25$, $MSE = 2500.00$, $p < .05$
7. Paired samples *t*-test: $t(14) = 3.65$, $r = .95$, $p < .01$
8. Independent groups *t*-test: $t(8) = 4.50$, $SE = 2.50$, $p < .05$ (two-tailed)
9. Chi-square test of independence: $\chi^2(1, N = 100) = 5.25$, $p < .05$
10. Chi-square test of independence: $\chi^2(4, N = 500) = 6.50$, $p < .01$

For #11 – 21, use the available information to calculate the needed sample sizes.

11. Independent groups *t*-test, Cohen's $d = .25$, $p = .01$ (two-tails), Power = .95
12. Independent groups *t*-test, Cohen's $d = .10$, $p = .05$ (one-tail), Power = .80
13. Paired samples *t*-test, Cohen's $d_z = .50$, $p = .001$ (two-tails), Power = .99
14. Paired samples *t*-test, Cohen's $d_z = .15$, $p = .005$ (one-tail), Power = .85
15. Oneway between subjects ANOVA, Cohen's $f = .05$, $p = .005$, $k = 4$, Power = .80
16. Oneway between subjects ANOVA, Cohen's $f = .15$, $p = .01$, $k = 3$, Power = .95
17. Oneway between subjects ANOVA, Cohen's $f = .30$, $p = .05$, $k = 5$, Power = .90
18. Pearson correlation, $r = .20$, $p = .0001$ (one-tail), Power = .95
19. Pearson correlation, $r = .40$, $p = .05$ (two-tails), Power = .80
20. Chi-square test for independence, Cohen's $w = .20$, $p = .05$, $df = 2$, Power = .85
21. Chi-square test for independence, Cohen's $w = .30$, $p = .001$, $df = 4$, Power = .95

For each of the following sets of results, calculate the appropriate effect size and then determine the number of subjects needed in a replication study with α = .05 and Power = .80, and a two-tailed test where appropriate.

22. Carter, Ferfuson and Hassin (2011) found that participants who were presented with a picture of the American flag reported a greater likelihood to vote for John McCain (M = 0.072, SD = 0.47) than people who were not presented with the American flag (M = -0.070, SD = 0.48), $t(181) = 2.02$, $p = .04$.

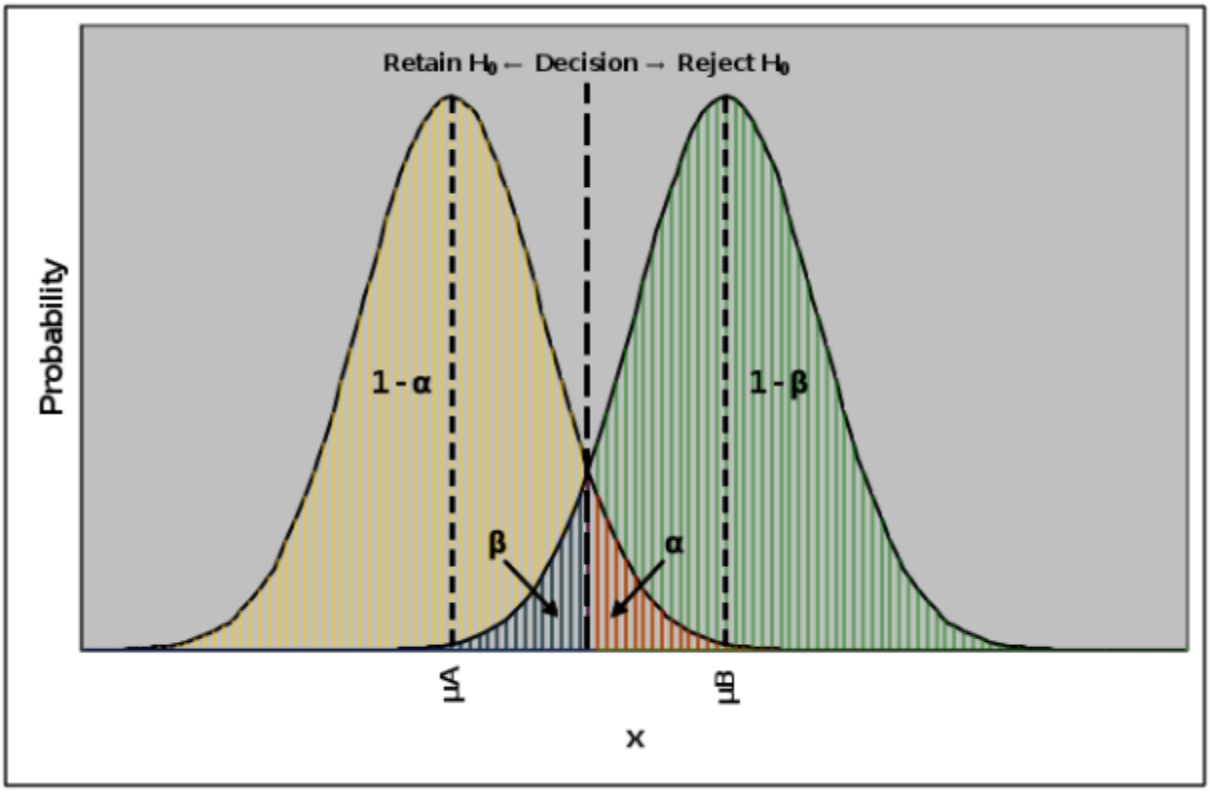
23. Kay. Moscovitch and Laurin (2010) had subjects unscramble each of 16 five-word sets. For half of the participants some of these words were related to randomness ("chance", "random") and for the other half the words were not related to randomness. Using a oneway, they found that subjects who were presented with the randomness-related words reported a stronger beliefs in the existence in supernatural sources of control than subjects who were not presented with randomness-related words, $F(1,33) = 4.47$, $p < .05$.

24. Tybur, Bryan, Magnan and Caldwell-Hooper (2011) observed that a group of subjects who were exposed to a noxious smelling compound ("Liquid ASS") reported a greater intention to use condoms than did a group of subjects who were not exposed to the noxious compound, $t(97) = 2.06$, $p < .05$.

Appendix D: Calculating Power

From the chapters on inferential testing, recall that all decisions with respect to rejecting or failing to reject a null hypothesis are based on probabilities; hence, there is always a degree of uncertainty that your decision about the null hypothesis was correct. More succinctly, your decision about the null could be wrong!

The alpha-level that you choose (α = .05, or less) is the probability that you are willing to make a Type I error. The probability of making a Type II error (β) is the probability you incorrectly retain the null hypothesis when it is actually false; that is, when you fail to reject the null when it is actually false. Finally, 1 - β is the **statistical power** that you have in your inferential test for correctly rejecting a false null hypothesis. The two overlapping distributions in the figure below demonstrate where α, β, 1 - α, and 1 - β occur:



Think of these as two distributions of scores measured from the same variable, where μ_A represents a distribution of scores where the null hypothesis is true, and μ_B represents a distribution of scores where the null hypothesis is false (and the alternate hypothesis is true). The point where the distributions overlap (the decision line) is the criterion on which you base your decision about a statistical test (the critical value/critical region). If there is sufficient evidence the null hypothesis is false, such that the evidence is greater than your criterion, you reject the null hypothesis. In contrast, if there is insufficient evidence, then you retain the null hypothesis. Basically, the decision whether to reject or retain the null comes down to your criterion, which is a critical value related to the α-level.

From the figure above, you can see a relationship between α, β, 1 - α, and 1 - β. The μ_A distribution is associated with both α and 1 - α, and the μ_B distribution is associated with β and 1 - β (power). It is important to remember that p[β + (1 - β)] = 1.00 and that p[α + (1 - α)] = 1.00; thus, [β + (1 - β)] = [α + (1 - α)]. Based on this, we can use any selected α level to determine the statistical power of an inferential test.

Computing the power of a statistical test makes use of z-scores. To calculate power you must know (i) the value of μ under the null hypothesis (the value to which you are comparing a sample mean), (ii) the value of the sample mean, (iii) the standard error of the mean, and (iv) the alpha-level. For the present example, assume that the alpha-level is .05 and we are using a two-tailed (non-directional) hypothesis. Also, assume that μ = 100, $\overline{X} = 105$, and $\sigma_{\overline{X}} = 2.5$

Calculating the power you have in a statistical test takes several steps that require calculating z-scores, looking up probabilities associated with z-scores from the z-tables (Appendix A, Table 1), and calculating raw scores from z-scores.

First, using the z-tables, find the z-score related to the chosen α-level (z_α). You will find that $z_\alpha = 1.96$.

Second, determine the value of a sample mean that would be associated with that z_α given μ, z_α, and the standard error of the mean. That is, what is the minimum mean value necessary to conclude that a sample mean is significantly different from a population mean? In the figure above, this is the value that is located at the decision line. This value is calculated as:

$$\overline{X}_\alpha = \mu + z_\alpha \sigma_{\overline{X}}$$
$$\overline{X}_\alpha = 100 + (1.96)(2.5)$$
$$\overline{X}_\alpha = 100 + 4.9$$
$$\overline{X}_\alpha = 104.9$$

Because both distributions lie along the same continuum, 104.9 would be found in both distributions; hence, it can be used as a reference point by both distributions.

Third, calculate the z-Score for the difference between 104.9 and the sample mean, which is labeled μ_B in the figure above:

$$z = \frac{\overline{X}_\alpha - \overline{X}}{\sigma_{\overline{X}}} = \frac{104.9 - 105}{2.5} = \frac{-0.1}{2.5} = -0.04$$

This is the z-score that is associated with your decision criterion in the alternate distribution; that is, the point where both distributions meet in the figure above. Look up this value in the z-tables. When you find this z-score, determine the probability of obtaining a z-score *less than or equal to* that value from Column 3. This will be the probability of observing a z-Score in the lower tail of the alternate distribution; thus, it is β, the probability of making a Type II error. In this example, β = .4840.

Finally, to determine the power of the test subtract β in the preceding step from 1.0; thus, the Power to detect a statistically significant result is 1 - β. In this example, the power (probability) to detect a statistically significant difference and to correctly reject the null hypothesis is 1.0 - .4840 = .5160. This is the probability that you will correctly reject the null hypothesis and accept the alternative hypothesis. This value ranges on a continuum from 0 to 1.00, with higher values indicating more statistical power.

Appendix F: Answers to Select Homework Questions

Note: Answers requiring verbal explanations or must be presented in your own words are not provided.

Chapter 1

1. a. Constant b. Variable c. Constant d. Constant e. Variable f. Variable

3. a. quantitative b. qualitative c. quantitative d. quantitative e. qualitative f. qualitative
g. qualitative h. qualitative i. quantitative

13. Independent variable: Dosage of drug, quantitative
Dependent variable: Brain activity, quantitative

15. Independent Variable: Whether subjects believed a rat was maze-bright or maze-dull, qualitative
Dependent Variable: Number of times the rat entered the correct section of the maze, quantitative

23. a. All undergraduates in the US
b. The 100 undergraduates who rated the potato-chips
c. Measurement, or raw data, or datum (any one is acceptable)
d. Descriptive statistic or a sample statistic (either is acceptable)

25. a. All statistic students.
b. The students in Dr. Smith's three statistics classes.
c. Measurement, or raw data, or datum (any one is acceptable)
d. Descriptive statistic or a sample statistic (either is acceptable)
e. Independent variable.
f. Ratio
g. Inferential statistics

Chapter 2

1. a. Ratio b. Nominal c. Ordinal d. Ordinal e. Interval f. Ratio

3. a. Ratio b. Nominal c. Ordinal d. Ratio e. Interval f. Ratio
g. Ordinal h. Ordinal i. Nominal

5. a. quantitative b. qualitative c. quantitative d. quantitative e. qualitative e. qualitative

7.

a. Number of Office Visits	f	rf	cf	crf
8	1	.05	20	1.0
7	5	.25	19	.95
6	8	.4	14	.7
5	2	.1	6	.3
4	2	.1	4	.2
3	2	.1	2	.1

b. .05 c. .95 d. 1 e. 0 f..5 g. 7
h. 3 i. 18 j. 4 k. 4

9.

College Major	f	rf	%
Undeclared (U)	29	.58	58 %
Psychology (P)	6	.12	12 %
Biology (B)	6	.12	12 %
English (E)	9	.18	18 %

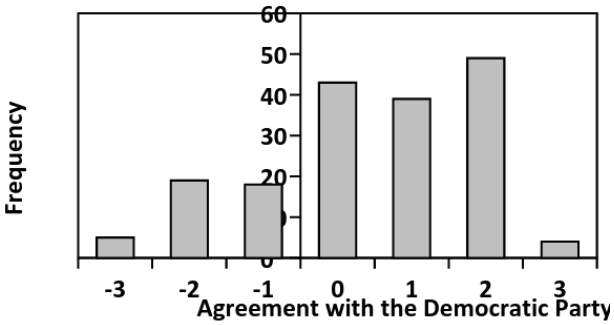
13.

Response	*%*	*f*	*rf*	*cf*	*crf*
4	35.0 %	700	.35	2000	1.00
3	15.0 %	300	.15	1300	.65
2	20.0 %	400	.20	1000	.50
1	30.0 %	600	.30	600	.30

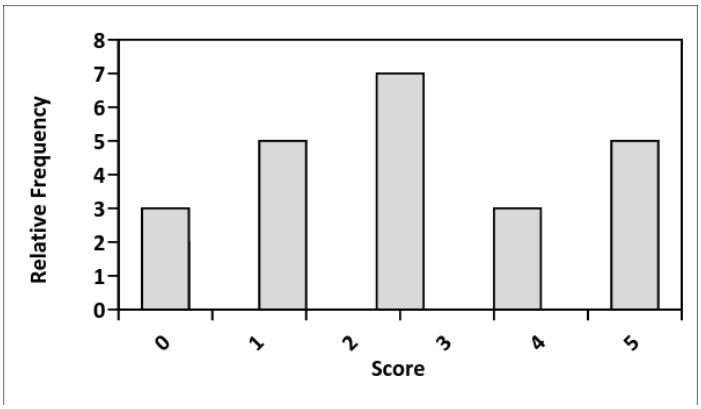
15. a. Interval Scale

b. Histogram, because the data is discrete and there are no values between any adjacent values.

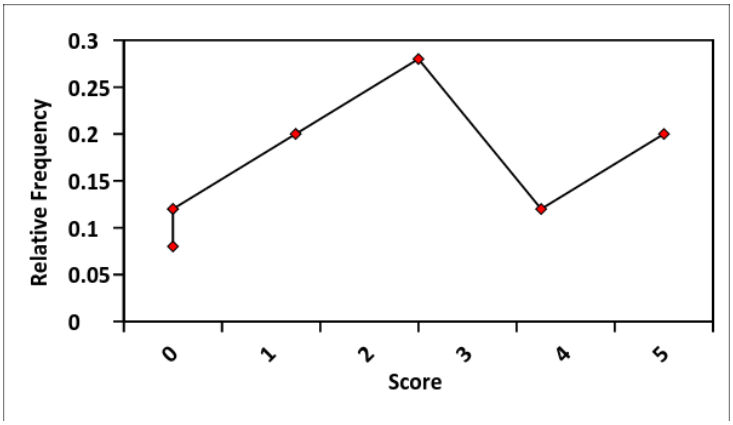
c.



17.

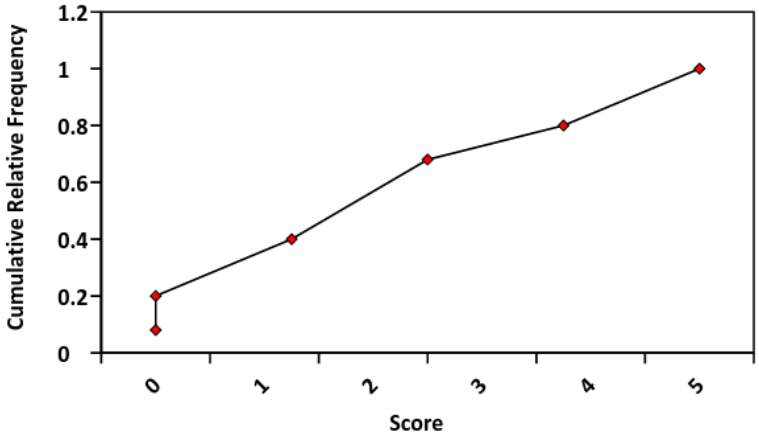


19.



The shape of the frequency polygon is the same as in #13.

21.



23.

Boys								Girls							
							6								
						9									
				5	8	4	1	2	8						
							0								
				3	3	6	1	3	9	4	1	3	8	2	5
							1								
2	2	2	6	9	2	0	1	2	5	4	2				
							2								
			4	4	3	6	1								
							3								
						4	1								
							4								

25.

a. What percentage of students scored higher than 4? 20%
b. What percentage of students scored a 3? 20%
c. What percentage of student scored less than 3? 60%

Chapter 3

1. a. 1.55 b. 2.714 c. 2.893 d. 3.65 e. 3.9 f. 5.375

3. a. 34% b. 58% c. 92% d. 16%

5. IQR = 5.294

Chapter 4

1. Mode, median, and mean

3. 5

7. a. $\Sigma X = 21$ b. $\Sigma(X - Y) = -9$ c. $\Sigma X^2 = 81$ d. $\Sigma(X)^2 = 81$ [Asks the same as c]
e. $(\Sigma X)^2 = 441$ f. $\Sigma(Y - X)^3 = 219$ g. $\Sigma XY = 90$ h. $\Sigma X - \Sigma Y = 21 - 30 = -9$

9. Mode = -1 Median = 0 Mean = -0.2

13. a. No single mode. b. 93.5 c. 92.8 d. Negatively skewed

15. a. 9 b. 8.5 c. 9 d. Negatively skewed

17. 48.28

21. The mean is a poorer measure of central tendency for Set A

23. Negatively skewed

Chapter 5

1. Sum of Squares, Variance, Standard Deviation

5. 49

7. a. 6.5 b. 28.5 c. 2.85 d. 1.688

11. a. 1125.6 b. 112.56 c. 10.609 d.125.067 e. 11.183

13. a. 22 b. 2.2 c. 1.483

15. Original Scores: $s^2 = 2$, s = 1.414 New Scores: $s^2 = 2$, s = 1.414

17. President = 5 Congress = 3

21. Correct Mean = 148 Correct Standard Deviation = 11

23. Mean = 24.5 SS = 26.5 Variance = 2.65 Standard Deviation = 1.628

Chapter 6

1. a. 0.240 b. -1.098 c. -3.658 d. 0 e. -1.013 f. 1.166

3. a. 16.75 b. 7 c. 7.75 d. 10

5. -1.225

7. Mean = 0, Standard Deviation = 1

9. a. 0.9750 b. 0.0250 c. 0.9505 d. 0.0495 e. 0.9162 f. 0.9162
g. 0.4772 h. 0.4772 i. 0.3433 j. 0.6541 k. 0.7850

11. a. 141.250 b. 100 c. 70 d. 77.5 e. 118.75 f. 110.5

13. a. 34300 b. 42900 c. 49600 d. 53300 e. 60800

15 a. 0.3707 b. 0.9525 c. 0.0475 d. 0.8413 e. 0.6293 f. 0.8164
g. 0.2120 h. 0.2286

Chapter 7

1. Difference between a sample statistic and a population parameter.

5. Number of scores that can vary in a sample.

7. Mean, standard deviation, and the shape of the sampling distribution.

9. The standard deviation of a sampling distribution of the mean.

11. The estimated standard error of the mean

13. The means are equal to the population mean.

15. $\sigma^2 = 3$ $\sigma = 1.732$ $\hat{\sigma}^2 = 3.333$ $\hat{\sigma} = 1.826$

17. $\sigma^2 = 1.8$ $\sigma = 1.342$ $\hat{\sigma}^2 = 2$ $\hat{\sigma} = 1.414$

19. 2.006

21. ± 5.33 LL = 60.67 UL = 71.33

23. a. 8 b. 6 c. 1.5 d. 1.225 e. 0.894
f. 6.428, 9.752 g. 0.548 h. 6.477, 6.523

27. 0.447

29. $\sigma_{\bar{X}} = 10$ 520 ± 19.6 500.4, 539.6

31. $\sigma_{\bar{X}} = 3.162$ 520 ± 8.158 511.842, 528.158

Chapter 8

3. a. 0.067 b. 0.167 c. 0.241 d. 0.31 e. 0.464

5. 0.4 7. 0.24 9. 0.06 11. 0.15 13. 0.1 15. 0.22

17. 0.36 19. 0.58 21. 0.4 23. 0.2 25. 0.552 27. 0.08

29. 0.8 31. 0.9 33. Yes 35. 0.5 37. 0.428 39. 0.4

41. 1 43. 0.348 45. 0.084 47. 0.228 49. 0.032 51. 0.6

53. 720 55. 10 57. 210

Chapter 9

1. a. 0.004 b. 0.112 c. 0.269 d. 0.032

3. Mean = 1670, Variance = 1391.11, Standard deviation = 37.298

5. Mean = 90, Variance = 36, Standard deviation = 6

7. Mean = 160, Variance = 32, Standard deviation = 5.657

9. p = 0.06, the student did not perform better than chance.

11. p = .123

13. a. 0.636 b. 54.4 c. 34.726 d. 5.893 e. z = 2.478, p = .0066

15. Mean = 20, Standard deviation = 4, z = 2.5, p = .0062

Chapter 10

5. a. 0.5 b. -3 c. .0027 d. Significant, Reject H_0, accept H_1.

7. a. H_0: μ = 100 H_1: μ > 100 b. 3 c. 1.333 d. .0918
e. Not significant, retain H_0, no decision on H_1.

9. a. H_0: μ = 13 H_1: μ > 13 b. 0.949 c. 0.823 d. .2061
e. Not significant, retain H_0, no decision on H_1. f. 0.3 g. 2.667
h. .0038 i. Yes. Significant, Reject H_0, accept H_1.

Chapter 11

3. a. .0600 b. .0182 c. .0194
d..0064 e. .0482 f. .0210

5. $\hat{\beta}_0 = 3.162$ $\beta = -0.949$ p = .2222 Not significant, retain H_0, no decision on H_1.

7. $\hat{\sigma}_{\bar{X}} = 2.4$ $t = -2.083$ p = .0484 Significant, reject H_0, accept H_1.

9. a. H_0: μ = 9 H_1: μ > 9 b. 1 c. 2.5 d. p. .0169
e. Significant, reject H_0, accept H_1.

Chapter 12

1. Experimental strategy involves manipulating an independent variable and the observational strategy involves measuring values that naturally exist in research participants.

3. IV: Noise condition (between-subjects). DV: Number of correct solutions.

5. A group not exposed to an experimental manipulation.

7. Variables that change as the levels of the independent variable change.

Chapter 13

3. a. .0096 b. .0224 c. .0214 d. .0439 e. .0442 f. 0073

5. $\hat{\sigma}^2_{pooled} = 4.809$ $\hat{\sigma}_{\bar{X}_1 - \bar{X}_2} = 0.922$

7. a. 15.53 b. 0.577 c. -17.953 or 17.953 d. .0001 e. Significant.

9. a. H_0: $\mu_{Audio} = \mu_{No\ Audio}$ H_1: $\mu_{Audio} \neq \mu_{No\ Audio}$ b. 1.875 c. 0.387 d. 2.067 or –2.067
e. 0.0450 f. Significant. Reject H_0, accept H_1.

11. a. 2.292 b. 0.307 c. -2.280 or 2.280 d. .0131
e. Significant. Reject H_0, accept H_1.

13. a. $M_M = 2.8$, $M_P = 1$, $SS_M = 1.3$, $SS_P = 0.5$ b. 0.225 c. 0.3
d. 6 e. .0039 f. Significant. Reject H_0, accept H_1.

15. a. H_0: $\mu_{Alcholo} = \mu_{Placebo}$ H_1: $\mu_{Alcohol} > \mu_{Placebo}$ b. $M_{Alcohol} = 2.5$ $M_{Placebo} = 1.75$
c. $SS_{Alcohol} = 0.128$ $SS_{Placebo} = 0.804$ d. 0.066
e. 0.126 f. 5.952 g. < .0001

17. a. .189 b. .860 c. .640 d. .798

19. $t(8) = 1.09$, $SE = 2.74$, $p = .1694$ (two-tails)

21. $t(8) = 6.00$, $SE = 0.30$, $p < .0039$ (two-tails)

Chapter 14

3. a. .0338 b. .0144 c. .0376 d. .0050

7. 0.408

9. a. 2 b. $SS_D = 10$, $\hat{\sigma}_D^2 = 2.5$, $\hat{\sigma}_D = 1.581$ c. 0.707 d. 2.829
e. .0474 f. Significant, Reject H_0, accept H_1. g. $\hat{\sigma}^2_{pooled} = 8.3$, $\hat{\sigma}^2_{pooled} = 2.3$
h. 2.302 i. 0.869 j. .321

11. a. H_0: $\mu_{Before} = \mu_{After}$; H_1: $\mu_{Before} < \mu_{After}$ b. $\hat{\sigma}_D^2 = 25.333$, $\hat{\sigma}_D = 5.033$
c. 1.592 d. 0.628 e. .2817 f. Not significant. Retain H_0, no decision on H_1
g. .042, small

13. a. 34 b. 1084 c. 21 d. 242

Chapter 15

9. a. .0224 b. .0112 c. .0442 d. .0221

11. a. $SS_X = 35$ $SS_Y = 22$ SCP = 22 b. $\hat{\sigma}_X = 1.784$ $\hat{\sigma}_Y = 1.414$ c. 2
d. 0.793 e. $r^2 = 0.629$ $1-r^2 = 0.371$ f. 4.13 g. .0022
h. Significant. Reject H_0, accept H_1. i. 0.981

13. a. H_0: $\rho = 0$, H_1: $\rho \neq 0$ b. -3.221 c. .0018
d. Significant. Reject H_0, accept H_1. e. 0.751

15. a. H_0: $\rho = 0$, H_1: $\rho < 0$ b. -3.462 c. .0006
d. Significant. Reject H_0, accept H_1.

17. a. $\sum D = 0$ $\sum D^2 = 90$ b. 0.455 c. 1.445
d. .1990 e. Not significant. Retain H_0, no decision on H_1.

19. a. $r' = 1.293$, $\rho' = 0.255$ b. 0.242 c. 4.289
d. .0001 e. Significant.

21. a. H_0: $\rho = -.50$, H_1: $\rho > -.50$ b. r' = -0.100, ρ' = -0.549 c. 0.102
d. 4.402 e. Significant. Reject H_0, accept H_1.

23. a. 0.395 b. 0.999 d. 0.637 e. 0.518

Chapter 16

5. If X changes by 1 unit, Y will change by 2.5 units.
If X changes by 3 units, Y will change by 7.5 units
If X changes by 8 unit, Y will change by 20 units.

7. a. -0.878 b. 8.390 c. Y' = 8.39 – 0.878X d. B = 0.488, H = 5.756, E = 3.122
e. $SS_{Regression} = 69.372$, $SS_{Residual} = 12.628$ f. 1.256

11. a. $SS_{Regression} = 74.34$, $SS_{Residual} = 9.66$ b. $df_{Regression} = 1$, $df_{Residual} = 8$, $df_{Total} = 9$
c. $MS_{Regression} = 74.34$, $MS_{Residual} = 1.208$ d. 61.54
e. .0133 f. Yes, significant.
g. 1.099 h. $SE(b_1) = 0.155$, $SE(b_0) = 0.712$
i. t = 7.871, $p = .0039$ j. t = -2.949, $p = .0184$

Chapter 17

9. a. .1313 b. .1016 c. .0388 d. .0036 e. .0258 f. .0135

11. $SS_B = 54$, $SS_W = 102.86$, $SS_T = 156.86$, $df_W = 20$, $MSW = 5.143$

13. $SS_B = 80$, $SS_W = 48$, $df_W = 12$, $df_T = 14$, $MS_B = 40$

15. 0.15 17. 0.18 19. 0.007

Chapter 18

1. a. $M_{Easy} = 7$, $M_{Mod} = 5$, $M_{High} = 3$, $M_{No} = 5$, $G = 5$
   b. $df_T = 19$, $df_B = 3$, $df_W = 16$
   c. $SS_W = 22$, $SS_T = 62$, $SS_B = 40$
   d. $MS_B = 13.333$, $MS_W = 1.375$
   e. 9.697
   f. .0007
   g. Significant. Reject H_0 and accept H_1.
   h. HSD = 2.122, Significant difference between Low Difficulty and High Difficulty.
   i. $\eta^2 = 0.645$, f = 1.348. Large effect.
   j. .997

3. a. 111 b. 177 c. 396 d. 280

5. a. $M_A = 7$, $M_B = 9$, $M_C = 9$, $M_D = 15$, $G = 10$
   b. $df_T = 23$, $df_B = 3$, $df_W = 21$
   c. $SS_W = 94$, $SS_T = 310$, $SS_B = 216$
   d. $MS_B = 72$, $MS_W = 4.476$
   e. 16.086
   f. .0003
   g. Significant. Reject H_0 and accept H_1.
   h. HSD = 4.337, Significant difference between A and D, B and D, C and D.
   i. $\eta^2 = 0.697$, f = 1.516. Large effect.
   j. .999

Chapter 19

1. Includes two or more independent variables

3. 9 groups, 2 independent variables

5. 16 groups; 4 independent variables

7. a. Yes b. Yes c. Yes

9. a. Yes b. Yes c. No

11.

a. Source of Variance	SS	df	MS	F
Between	117	3	39	3.9
Main Effect of X	100	1	100	10
Main Effect of Y	2	1	2	0.5

Interaction (X x Y)	15	1	15	1.5
Within Groups	360	36	10	---
Total	477	39	---	---

b. Source of Variance	SS	df	MS	F
Between	151	3	50.333	2.013
Variable X	1	1	1	0.04
Variable Y	50	1	50	2
Interaction	100	1	100	4
Within	1900	76	25	---
Total	2051	79	---	---

13. a. Main Effect of X: $p = .0036$ Main Effect of Y: $p = .3253$ Interaction: $p = .2302$
b. Main Effect of X: $p = .3213$ Main Effect of Y: $p = .1625$ Interaction: $p = .500$

15. a. $M_{J1/K1} = 9$, $M_{J1/K2} = 3$, $M_{J2/k1} = 2$, $M_{J2/K2} = 8$, $M_{J1} = 6$, $M_{J2} = 5$, $M_{K1} = 5.5$, $M_{K2} = 5.5$, $G = 5.5$
b. $df_T = 11$, $df_W = 8$, $df_B = 3$, $df_J = 1$, $df_K = 1$, $df_{JxK} = = 1$
c. $SS_T = 119$, $SS_W = 8$, $SS_B = 111$, $SS_J = 3$, $SS_K = 0$, $SS_{JxK} = 108$,
d. $MS_B = 37$, $MS_{Within} = 1$, $MS_J = 3$, $MS_K = 0$, $MS_{Interaction} = 108$
e. $F_J = 3$, $F_K = 0$, $F_{Interaction} = 108$
f. $p_J = .1215$, $p_K = .3466$, $p_{Interaction} = .0133$
g. Neither main effect is statistically significant; however, the interaction is statistically significant.

Chapter 20

3. Observed frequencies are the actual numbers. Expected frequencies are expected under a null hypothesis.

5. a. .0339 b. .1069 c. .0514 d. .1091
e. .0614 f. .0174 g. .0937

7. a. 10 b. 347.4 c. .0000 (or < .0001) d. Significant. Reject H_0, accept H_1.

9. a. Democrat-City = 104, Democrat-Suburbs = 104, Democrat-Country = 52,
Republican-City = 56, Republican-Suburbs = 56, Republican-Country = 23,
Independent-City = 40, Independent-Suburbs = 40, Independent-Country = 20
b. 26.374 c. .0000 (or < .0001) d. Significant. Reject H_0, accept H_1.
e. 0.224 f. 0.230 g. 0.698

11. a. Natural Sciences = 20, Business School = 25, Social Sciences = 35, Arts and Humanities = 20
b. 27.478 c. .0000 (or < .0001) d. Significant. Reject H_0, accept H_1.

13. a. 15.425 b. .0001 c. Significant. Reject H_0, accept H_1. d. 0.507

15. a. 964 b. 241 c. 1545 d. 299

Chapter 21

1. d = 0.73

3. $\hat{\sigma}_1^2 = 25$, $\hat{\sigma}_2^2 = 81$, $\hat{\sigma}_{\square\square\square\square\square\square} = 7.280$, d = 0.412

5. $f = 0.345$ or $\square^2 = 0.108$

7. d = 0.307

9. C = 0.224, $w = 0.230$

11. 1144

13. 132

15. 6940

17. 180

19. 44

21. 353

23. f = 0.366, n = 62

References

Bem, D. J. (2011). Feeling the Future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 407-425.

Carter, T. J., Ferguson, M. J., & Hassin, R. R. (2011). A single exposure to the American flag shifts support toward Reublicanism up to 8 months later. *Psychological Science, 22*, 1011-1018.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1*, 170-177.

French, C. (2012, March 15). Precognition studies and the curse of the failed replications. *The Guardian*. Retrieved from http://www.guardian.co.uk/science/2012/mar/15/precognition-studies-curse-failed-replications

Kay, A. C., Moscovitch, D. A., & Laurin, K. (2010). Randomness, attributions of arousal, and belief in God. *Psychological Science, 21*, 216-218.

Liptak, A. (2012, June 7). Approval rating for justices hits just 44% in new poll. *The New York Times*. [Website] Retrieved from http://www.nytimes.com/2012/06/08/us/politics/44-percent-of-americans-approve-of-supreme-court-in-new-poll.html?pagewanted=all

Newport, F. (2009, February 11). On Darwin's birthday, only 4 in 10 believe in evolution. Retrieved from http://www.gallup.com/poll/114544/Darwin-Birthday-Believe-Evolution.aspx

Ritchie SJ, Wiseman R, French CC (2012) Failing the Future: Three Unsuccessful Attempts to Replicate Bem's 'Retroactive Facilitation of Recall' Effect. *PLoS ONE* 7, e33423. doi:10.1371/journal.pone.0033423

Rosnow, R. L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods, 1*, 331-340.

Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science, 11*, 446-453.

Shah, J., & Christopher, N. (2002), Can shoe size predict penile length? *BJU International, 90*, 586–587. doi: 10.1046/j.1464-410X.2002.02974.x

Tybur, J. M., Bryan, A. D., Mangan, R. E., & Caldwell Hooper, A. E. (2011). Smells like safe sex: Olfactory pathogen primes increase intentions to use condoms. *Psychological Science, 22*, 478-480.

The Following were used as References for Statistical Formulas

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd Ed.)*. Hillsdale, NJ: Lawrence Erlbaum.

Hays, W. L. (1994). *Statistics (5th Ed.)*. Belmont, CA: Thompson-Wadsworth.

Heiman, G. W. (2006). *Basic statistics for the behavioral sciences (5th Ed.).*Boston, Houghton Mifflin.

Keppel, G. (1991). *Design and analysis: A researcher's handbook (3rd Ed.)*. Upper Saddle River, NJ: Prentice Hall.

Jaccard, J., & Becker, M. A. (2010). *Statistics for the behavioral sciences (5th Ed.)*. Belmont, CA: Wadsworth Cengage.

Myers, J. L., & Well, A. D. (2003). *Research design and statistical analysis (2nd Ed.)*. Mahwah, NJ: Lawrence Erlbaum.

Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis (3rd Ed.).* Boston: McGraw-Hill.

Timhane, A. C., & Dunlop, D. D. (2000). *Statistics and data analysis: From the elementary to the intermediate.* Upper Saddle River, NJ: Prentice Hall.